Pilotstudie zu einer Automatisierung der Zeitungsdokumentation des Année Politique Suisse / Jahrbuch Schweizerische Politik

Wüest, Bruno

Other titles: Pilot study on the automated newspaper classification at the Yearbook of Swiss Politics

Pilotstudie

zu einer Automatisierung der Zeitungsdokumentation des Année Politique Suisse / Jahrbuch Schweizerische Politik

Bruno Wueest*

Universität Zürich[UZH]

swiss online politics

Last update: 12. Januar 2014

Dieser Bericht präsentiert und erläutert die Resultate der Pilotstudie zur computergestützten Zeitungsdokumentation des Année Politique Suisse / Jahrbuch Schweizer Politik (APS). Zunächst werden die Erkenntnisse der Pilotstudie für die Automatisierung der Selektion und Klassifikation der Zeitungsanalysen diskutiert. Es zeigt sich, dass eine Automatisierung bei den ersten Schritten der APS-Zeitungsdokumentation (Selektion relevanter Zeitungsartikel und Erhebung der Zeitungsartikelangaben wie z.B. das Publikationsdatum) sehr hilfreich sein kann. Aber auch für eine grobe ertse Klassifikation wurden gute Resultate erzielt, die eine Teilautomatisierung der Themeneinteilung Nahe legen. Aus den Resultaten der Pilotstudie werden in der Folge konkrete Vorschläge für eine Reorganisation der Codierungsprozesse abgeleitet. Bei verstärktem Einsatz von Automatisierungen würden sich konkret die Entwicklung einer Codierapplikation, die Reorganisation der Abläufe der APS-Zeitungsdokumentation sowie ein verstärktes Engagement für die Sicherung der Reliabilität aufdrängen.

Inhaltsverzeichnis

*Oberassistent in Lehre und Forschung, Institut für Politikwissenschaft, Universität Zürich. Leiter des Teilprojekts *Swiss Online Politics*. Email: `wueest@ipz.uzh.ch`; Web: `http://www.bruno-wueest.ch/`

1 Pilotstudie

Die Pilotstudie hat sich mit den in Abbildung 1 schematisch dargestellten Arbeitschritten der APS-Zeitungsdokumentation befasst. Angefangen wird bei relativ einfachen Verfahren wie der Extraktion von Metadaten[1] wird während der Studie zu immer komplexeren Aufgaben wie dem immer feineren Klassifizieren der Zeitungsartikel in die APS-Themenkategorien übergegangen. Die grauen Schattierungen illustrieren die Erwartungen an die Evaluationen in Bezug auf den Schwierigkeitsgrad: Von der Extraktion der Metadaten wird erwartet, dass diese mit praktisch 100-prozentiger Genauigkeit möglich ist, weil dies eine einfache Identifikation von hoch standardisierter Information ist. Bei der Selektion relevanter Artikel handelt es sich bereits um eine erste Klassifikation, die naturgemäss unsicherer ist, weil sie mit statistischen Verfahren berechnet werden muss. Allerdings ist es eine binäre Klassifikation (Einteilung der Artikel in *relevant/nicht relevant*), was die Aufgabe erheblich vereinfacht. Anschliessend wird die grobe Themenklassifikation auf der obersten Stufe des APS-Klassifikationsschemas versucht (8 Themenkategorien sowie eine Kategorie für Gegenstände und Träger der Politik). Diese Aufgabe ist schwieriger aber erwartungsgemäss machbar, weil diese Themenkategorien noch relativ klar voneinander abgrenzbar sind. Eine feinere Klassifikation in die drei- und vierstelligen APS-Themenkategorien ist dann dem experimentellen Bereich zuzuordnen und es ist ungewiss, ob hier eine ausreichende Qualität erreicht wird.

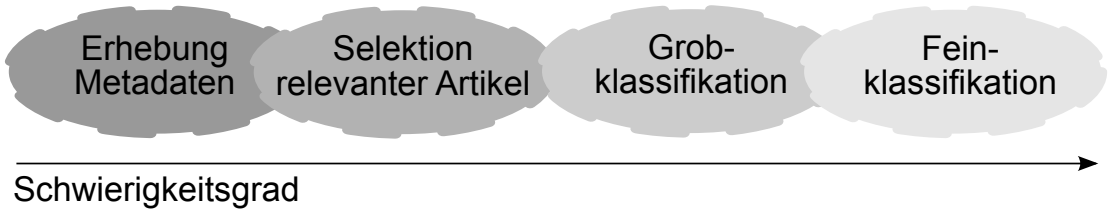

Abbildung 1: Aufbau der Pilot-Studie und Schwierigkeitsgrad für die Automatisierung

Für alle Tests der Automatisierung der Arbeitsschritte wurde die gesammten manuell erarbeiteten Klassifizierungen des APS vom Januar bis November 2013 mit den folgenden Einschränkungen berücksichtigt: a) eine Zeitung muss im APS-Standardsample und nicht in der Spezialkategorie zu den Abstimmungskampagnen aufgeführt sein, b) der Zeitungstitel muss in der Datenbank des Schweizerische Mediendienstes (SMD) indexiert sein – nur über den SMD konnte eine systematisches Beschaffung der Artikeln erfolgen (siehe Tabelle 10 für eine Aufstellung der Verfügbarkeiten), und c) die Zeitung muss in Deutsch oder Französisch publizieren, weil für italienisch keine sprachspezifische Software eingesetzt werden konnte. Diese Auswahl ergibt insgesamt 26'422 Zeitungsartikel, 20'042 in deutscher und 6'380 in französischer Sprache. Aus dieser Grundgesamtheit wurde aus Machbarkeitsgründen eine Zufallsstichprobe von 10'000 Artikeln gezogen, mit welcher die Selektions- und Klassifizierungstests durchgeführt wurden. Tabelle 1 zeigt die Verteilung der Artikel über die beiden Sprachen und einzelnen Titel an. Weil es einzelne Probleme beim Download der Artikel aus der SMD-Datenbank gab, ist die Gesamtzahl der Artikel leicht unter dem Richtwert (9'921 statt 10'000 Artikel). Diese Fehler sind aber in keinster Weise systematisch und die Stichprobe kann somit als repräsentativ für die APS-Zeitungsdokumentation gelten. Die Zeitungsartikel dieser Auswahl sind im Dropbox-Ordner vorhanden, welcher dem APS gleichzeitig mit der Fertigstellung dieses Berichts zur Verfügung gestellt wurde.

[1] Als Metadaten werden hier Informationen zu den einzelnen Zeitungsartikeln verstanden, d.h. der Zeitungtitel, das Publikationsdatum, die Seite und der Titel eines Artikels.

Tabelle 1: In der Evaluation berücksichtigte Zeitungsartikel (Zufallsauswahl)

Klassifizierung	
Aargauer Zeitung	850
Basler Zeitung	863
Blick	267
Bund	156
Berner Zeitung	375
Basellandschaftliche Zeitung	553
L'Express	282
24 Heures	204
La Liberté	353
Neue Luzerner Zeitung	222
Nouvelliste	212
Neue Zürcher Zeitung	1'558
Sonntagsblick	153
St. Galler Tagblatt	472
Südostschweiz	322
Solothurner Zeitung	261
Tages-Anzeiger	730
Sonntags-Zeitung	213
Tribune de Genève	802
Le Matin	127
Le Temps	426
Thurgauer Zeitung	330
Walliser Bote	33
Weltwoche	143
Wochenzeitung	11
Deutschsprachige Titel	7'513
Französischsprachige Titel	2'407
Total	9'921

Insgesamt ist die Verteilung der Artikel über die Zeitungstitel sehr gut, denn von den meisten Zeitungen sind mehr als 200 Artikel vorhanden. Von einigen Zeitungen, vor allem von den Wochenzeitungen, hat es allerdings etwas wenige Artikel, was jedoch auch der tatsächlichen Verteilung in der Grundgesamtheit der APS-Zeitungsdokomentation entspricht. Damit diese wenigen Artikel nicht zum Problem während der Evaluation werden, werden die Selektion und Klassifikationen zunächst generell für alle Artikel der beiden Sprachen durchgeführt. Zusätzlich werden die Berechnungen auch für die vier wichtigsten Zeitungen (Titel mit über 800 Artikeln: Aargauer Zeitung, Basler Zeitung, Neue Zürcher Zeitung und Tribune de Genève) separat durchgeführt. Somit wird eine solide Überprüfung der Leistung der Selektions- und Klassifikationsverfahren generall aber auch auf dem Niveau der einzelnen Titel ermöglicht.

Die vorliegende Pilotstudie bietet eine systematische Evaluation der Qualität der Automatisierungen und soll somit eine Entscheidungsgrundlage bieten, welche Automatisierungen für allfällige im Alltag einsatzfähige Weiterentwicklungen sinnvoll sind. Zudem werden alle Schritte der Studie zunächst anhand Ausschnitten der eigens geschriebenen Software erläutert. Nach der Diskussion der Resultate der Evaluationen werden schliesslich konkrete Handlungsanweisungen abgeleitet, welche auf die weiteren Schritte zur Reorganisation der APS-Zeitungsdokumentation abgeleitet. Die Automatisierungen der Selektion und Klassifikation wurden einzeln entwockelt und auf ihre Tauglichkeit getestet. Das heisst es existiert kein integriertes Softwarepaket, welches einen Alltagseinsatz erlauben würde. Die Entwicklung einer umfassenden Softwarelösung für das APS würde dann aktuell werden, wenn ein Grundsatzentscheid zugunsten weiterer Automatisierungen gefällt wird.

1.1 Verwendete Software

Die hier dokumentierte Pilotstudie wurde mit vollständig in *R* geschriebener Software durchgeführt (vgl. `http://www.r-project.org/`). *R* ist eine sehr versatlie Software-Umgebung und ermöglicht die

Integration der für eine Automatisierung der APS-Zeitungsdokumentation notwendigen Technologien wie Datenbanksteuerung, Webserver-Applikationen, Webscraping-Funktionen und Textanalyse-Programmen. Die Software-Umgebung *R* ist nicht nur deshalb ideal für die in der Folge präsentierten computergestützten Klassifikationen und Codierungsprozesse, sondern auch weil sie frei verfügbar (kostenlos und Open Source) ist sowie auf allen Betriebssystemen läuft. In den folgenden Abschnitten werden die spezifisch für die APS-Dokumentation geschriebenen Skripte präsentiert und erläutert. Die wichtigsten dabei verwendeten *R*-Programmbibliotheken sind *tm* (Feinerer, Hornik and Meyer, 2008) und RTextTools (Jurka et al., 2013), ersteres für allgemeine Operationen an Texten und letzteres speziell für die Textklassifikationen. Alle Skripte sind im Dropbox-Ordner der Pilotstudie auffindbar.

In der Prämbel aller Skripte sind einige gemeinsame Einstellungen notwendig, die zunächste kurz beschrieben werden. Es handelt sich um 1) die Bereinigung des *R*-Arbeitsspeichers, 2) die Definition des Arbeitsverzeichnis und 3) die Korrektur irreführender Eigenheiten von *R* (keine kategorialen Variablen (*factors*) aus Textdaten generieren und Rundungsfehler bei mathematischen Operationen konstant halten).

1. Skript-Auszug: Präambel

```
# Voreinstellungen
rm(list=ls(all=TRUE)) # alles im Arbeitsspeicher entfernen um keine störenden
# Einflüsse zu haben
setwd(''~/path/to/working/dir'') # den Pfad zum Arbeitsverzeichnis setzen
options(stringsAsFactors=F) # die automatische Konvertierung von String-Variablen in
# Faktoren verunmöglichen
set.seed(12345) # Zufallszahl fixieren um die Replizierbarkeit sicherzustellen
```

1.1.1 Extraktion der Metadaten und des Rohtexts der Artikel

Als erster Schritt nach dem Herunterladen der Zeitungsartikel können die Metadaten erhoben werden. Die vom SMD heruntergeladenene Dateien sind genug standardisiert, dass sie eine automatisierte Codierung des Publikationsdatums, Zeitungsnamens sowie Artikeltitels erlauben. In den Abbildungen 2 und 3 ist der Kopf eines Zeitungsartikels im SMD-HTML-dargestellt. In der oberen Hälfte wird die Version eines SMD-Zeitungs-Artikels gezeigt, welche normalerweise in einem Browser ersichtlich ist. Der dhinterliegende Quellcode, der dann auch analysiert und codiert werden kann, ist in der unteren Hälfte von Abbildung 2 dargestellt. Die Angaben zur Zeitung, dem Publikationsdatum sowie der Seite (Aargauer Zeitung ...) beispielsweise sind in ein HTML-Tag mit dem Namen * * eingebettet. Dieses Tag kann mit dem weiter unten erklärten Software-Skript angesteuert werden, womit diese Metadaten codiert werden können.

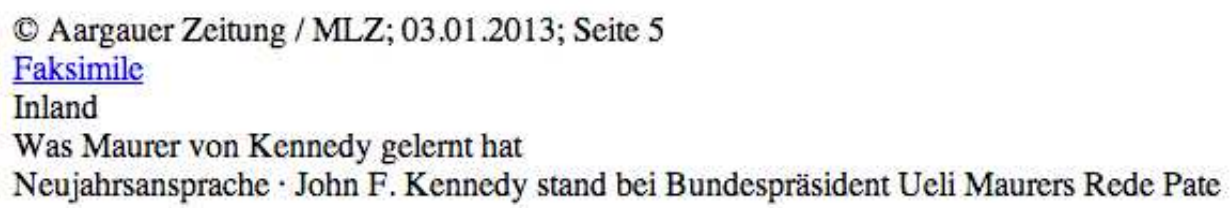

Abbildung 2: Originalansicht des Kopfs eines HTML-Zeitungsartikels

```
<!--?xml version="1.0" encoding="UTF-8"?-->
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="de" xml:lang="de"><head>
      <meta content="text/html; charset=UTF-8" http-equiv="Content-Type" />
      <link media="screen, print" type="text/css" href="/css/smd.css" rel="StyleSheet" />
      <title>SMD Dokument</title>
   </head>
   <body>
      <div style="float: right; ">
         <a href="http://www.smd.ch" name="top">
            <img src="/utils/logo-300.gif" alt="SMD Schweizer Mediendantenbank AG" style="border-width: 0; " />
         </a>
      </div>
      <div class="topHeader">
         <a name="JM20130103000499833">© Aargauer Zeitung / MLZ; 03.01.2013; Seite 5</a>
      </div>
      <div class="docLinks">
         <a href="/SmdDocuments/?userInterface=SMDDocuments&amp;aktion=protectedDocumentsDownload&amp;view=PDFPageScr
      </div>
      <div class="titleSection">
         <div class="RE"> Inland</div>
         <div class="HT">Was Maurer von Kennedy gelernt hat</div>
         <div class="UT">Neujahrsansprache ·  John F. Kennedy stand bei Bundespräsident Ueli Maurers Rede Pate</div>
      </div>
      <div class="body">
```

Abbildung 3: Quellcode des Kopfs eines HTML-Zeitungsartikels

Die ersten Zeilen im Softwareskript welches solche Metadaten codiert sind die folgenden (das dazugehörige Program im Dropbox-Ordner heisst *APS_Vorbereitung.r*):

2. Skript-Auszug: Vorbreitung I

```
# Notwendige Packete laden
# ..wenn nötig mit ''install.packages(c(<Liste mit Namen der Packete>))'' installieren
library(tm) # Vielseitiges Textberarbeitungstool
library(XML) # Fuer XPath
library(stringr) # Regular Expressions

# Pfade zu den heruntergeladenen Dateien definieren
directory <- ''1_html/''
docnames <- list.files(directory)
```

Zunächst müssen drei Pakete geladen werden: *tm* um generell den Umgang mit Textdaten zu ermöglichen, *XML* um mit Webfomraten (HTML, XML etc.) arbeiten zu können und *stringr* um Reguläre Ausdrücke einzusetzen. Reguläre Ausdrücke sind eine in der maschinellen Textbearbeitung weit verbreitete Technologie, um mit Sonderzeichen generelle Muster in Texten zu erkennen (z.B. werden oft Email-Adressen oder Telefonnummern mit Regulären Ausdrücken erkannt, weil sie eine standardisierte Struktur aufweisen)[2]. Schliesslich wird der Pfad zum Verzeichnis, wo die Artikel abgespeichert wurden, definiert sowie eine Liste der Dateinamen der Artikel generiert. Anschliessend können mit dem folgenden Loop die Metadaten erhoben werden:

3. Skript-Auszug: Vorbreitung II

```
# Einen Dataframe als 'Container' für die Resultate definieren
metaDF <- data.frame()

# Laufzahl für die Identifikation definieren
no <- 1

# Loop starten
for (name in docnames) {
```

[2]Für genauere Erklärungen hierzu siehe `http://www.regular-expressions.info/`.

```
  # Dokument parsen und Liste mit Metadaten extrahieren
  parse <- htmlParse(paste0(''1_html/'',name), encoding=''UTF-8'')
  meta <- xpathSApply(parse, ''//div[@class='topHeader']'', xmlValue) # die Metadaten
  # sind im Abschnitt, der mit ''topHeader'' gekennzeichnet ist
  meta <- gsub(''[[:space:]]'', '', meta)
  meta <- unlist(str_split(meta, '';''))

  outlet <- gsub(''[^[:alnum:] ]'', '', meta[1]) # Name der Zeitung extrahieren

  date <- gsub(''[[:punct:]]'', '-', meta[2]) # Datum extrahieren

  page <- grep(''^Seite'', meta, value = T) # Seitennummer extrahieren
  page <- gsub(''[^[:digit:] ]'', '', page)

  title <- xpathSApply(parse, ''//div[@class='HT']'', xmlValue) # Kurztitel extrahieren
  title <- gsub(''^\\s+|\\s+$'', '', title) # Leerzeichen zu Beginn und am Ende entfernen

  # Zeile mit Metadaten in Datenframe einfügen
  metaDF <- rbind(metaDF, cbind(name, no, outlet, date, page, title))

  }

# Metadaten herausschreiben
write.table(metaDF, file = ''meta.tsv'', quote = F, sep = ''\t'', fileEncoding = ''UTF-8'',
    row.names = F)
```

Ein *data.frame*, der zunächst initialisiert wird, ist ein häufiges Datenformat in *R*, welches eine tabellenartige Speicherung erlaubt. Anschliessend wird ein *for*-Loop aufgerufen, der die Metadaten extrahiert und in den *data.frame* füllt. Nacheinander werden mit *XPath* und Regulären Ausdrücken (Funktion *gsub*) Zeitungsname (*outlet*), Publikationsdatum (*page*), Seitennummer (*page*) und Titel (*title*) identifiziert. Schliesslich werden alle Metadaten in einem Spreadsheet-File mit dem Namen *meta.tsv* gespeichert. Das automatisiert generierte Metadatenfile ist im Dropbox-Ordner auffindbar. Die Reliabilität dieser Extraktionsmethoden wird unter Abschnitt 1.2 präsentiert.

Nach den Metadaten gehört auch die Extraktion des eigentlichen Texts der Zeitungsartikel zur Vorbereitung. Normalerweise enthalten Webformate wie HTML noch eine Fülle von Informationen, z.B. Schriftformate oder Bilddateien. Zwischen all diesen weiterern Informationen muss der Rohtext der Artikel herausgeholt werden, was mit dem unten aufgeführten Code geleistet werden kann.[3]

4. Skript-Auszug: Vorbreitung III

```
  # Identifikationsname für die Speicherung zusammensetzen
  ID <- paste(outlet, date, no, ''txt'', sep = ''.'')

  # Gesamten Titel extrahieren
  title <- xpathSApply(parse, ''//div[@class='titleSection']'', xmlValue)
  title <- gsub(''^\\s+|\\s+$'', '', title)
  title <- gsub(''\n\\s+'', '\n', title)

  # Textkörper extrahieren
  text <- xpathSApply(parse, ''//div[@class='body']'', xmlValue)
  text <- gsub(''^\\s+|\\s+$'', '', text)
  text <- gsub(''\n\\s+'', '\n', text)

  # Alles zusammenfügen für die Rohtextfiles
  output <- paste(title, text, '', sep = ''\n'')
  output <- gsub('-', ' ', output)
  output <- gsub(''[^[:alnum:][:space:]]'', '', output)
```

[3] Im fertigen Programm sind diese Passagen in den im 3. Skript-Auszug aufgeführten Loop Stelle integriert.

```
# Titel und Text unter dem Identifikationsnamen speichern
cat(output, file = (con <- file(paste0('3_raw/', ID), 'w', encoding = 'UTF-8')));
      close(con)

no <- no + 1 # Laufzahl erhöhen
```

Diese Skript-Passagen definieren zuerst eine Identifikationsname bestehend aus dem Namen der Zeitung, dem Publikationsdatum und einer Laufzahl. Somit bleiben die Artikel für alle weiteren Schritte eindeutig identifizierbar. Anschliessen wird der gesamte Text – enthalten in der Titelsektion und dem Textkörper (gekennzeichnet durch das HTML-Tag *body*) – extrahiert, zu einem Rohtext-objekt zusammengefügt und mit der Identifikationssequenz als Dateinamen abgespeichert. In einer wirklich einsatzfähigen Codierapplikation müssten die Texte und Metadaten nun in eine Datenbank eingespiesen werden, weil dort selbst grosse Mengen an Artikeln effizient gespeichert und schnell abgerufen werden können.

1.1.2 Vorverarbeitung der Artikel

Ein essenzieller Schritt für eine automatisierte Klassifikation ist die Bereitstellung verschiedener Versionen der Texte, die zu klassifizieren sind. Textklassifikation ist prinzipiell eine statistische Aufgabe, weshalb verschiedene Verfahren getestet werden können, welche die verfügbare Information in den Texten auf die wesentlichen Teile zu beschränken versuchen. Im Detail werden die folgenden Informationsreduktionsverfahren zur Vorbereitung der Klassifikation angewendet:

Stopping Ein in der Webtechnologie sehr verbreitetes Verfahren (z.B. für die Optimierung von Suchmaschinen). Hier wird in Texten eine Liste aus Worten ausgeschlossen, welche nicht für alle Aufgaben als gehaltvoll für den Inhalt eines Textes angesehen werden. Dies sind meistens kürzere Worte wie Artikel oder Pronomen sowie Hilfsverben. Für die Untersuchung wurden die foldenen Standard-Stoppwortlisten von *tm* verwendet:

Französisch *au, aux, avec, ce, ces, dans, de, des, du, elle, en, et, eux, il, je, la, le, leur, lui, ma, mais, me, même, mes, moi, mon, ne, nos, notre, nous, on, ou, par, pas, pour, qu, que, qui, sa, se, ses, son, sur, ta, te, tes, toi, ton, tu, un, une, vos, votre, vous, c, d, j, l, à, m, n, s, t, y, été, étée, étées, étés, étant, suis, es, est, sommes, êtes, sont, serai, seras, sera, serons, serez, seront, serais, serait, serions, seriez, seraient, étais, était, étions, étiez, étaient, fus, fut, fûmes, fûtes, furent, sois, soit, soyons, soyez, soient, fusse, fusses, fût, fussions, fussiez, fussent, ayant, eu, eue, eues, eus, ai, as, avons, avez, ont, aurai, auras, aura, aurons, aurez, auront, aurais, aurait, aurions, auriez, auraient, avais, avait, avions, aviez, avaient, eut, eûmes, eûtes, eurent, aie, aies, ait, ayons, ayez, aient, eusse, eusses, eût, eussions, eussiez, eussent, ceci, cela, celà, cet, cette, ici, ils, les, leurs, quel, quels, quelle, quelles, sans, soi*

Deutsch *aber, alle, allem, allen, aller, alles, als, also, am, an, ander, andere, anderem, anderen, anderer, anderes, anderm, andern, anderr, anders, auch, auf, aus, bei, bin, bis, bist, da, damit, dann, der, den, des, dem, die, das, daß, derselbe, derselben, denselben, desselben, demselben, dieselbe, dieselben, dasselbe, dazu, dein, deine, deinem, deinen, deiner, deines, denn, derer, dessen, dich, dir, du, dies, diese, diesem, diesen, dieser, dieses, doch, dort, durch, ein, eine, einem, einen, einer, eines, einig, einige, einigem, einigen, einiger, einiges, einmal, er, ihn, ihm, es, etwas, euer, eure, eurem, euren, eurer, eures, für, gegen, gewesen, hab, habe, haben, hat, hatte, hatten, hier, hin, hinter, ich, mich, mir, ihr, ihre, ihrem, ihren, ihrer, ihres, euch, im, in, indem, ins, ist, jede, jedem, jeden, jeder, jedes, jene, jenem, jenen, jener, jenes, jetzt, kann, kein, keine, keinem, keinen, keiner, keines, können, könnte, machen, man, manche, manchem, manchen, mancher, manches, mein,*

meine, meinem, meinen, meiner, meines, mit, muss, musste, nach, nicht, nichts, noch, nun, nur, ob, oder, ohne, sehr, sein, seine, seinem, seinen, seiner, seines, selbst, sich, sie, ihnen, sind, so, solche, solchem, solchen, solcher, solches, soll, sollte, sondern, sonst, über, um, und, uns, unse, unsem, unsen, unser, unses, unter, viel, vom, von, vor, während, war, waren, warst, was, weg, weil, weiter, welche, welchem, welchen, welcher, welches, wenn, werde, werden, wie, wieder, will, wir, wird, wirst, wo, wollen, wollte, würde, würden, zu, zum, zur, zwar, zwischen

Mindestwortlänge Ähnlich zum Stopping kann angenommen werden, dass besonders kurze Worte nicht relevant sind für eine Klassifikationsaufgabe. In dieser Pilotstudie wird getestet, ob der Anschluss von Worten mit weniger als 3 Buchstaben bessere Resultate erzielt.

Zahlen Es kann versucht werden, durch einen Ausschluss von Zahlen eine Verbesserung der Klassifikation zu erreichen. In der vorliegenden Studie kann erwartet werden, dass Zahlen nicht entscheidend sind für die Unterscheidung verschiedener politischer Themen wie z.B. Sozial- und Aussenpolitik und somit weggelassen werden können..

Stemming Alle konjugierten und deklinierten Worte in einem Text können auf ihre Stammform zurückgesetzt werden. Hier wird dazu ein einfaches Verfahren verwendet, welches die Wortendungen löscht.

Kleinbuchstaben Alle Worte werden nur in Kleinschreibung in die Berechnungen aufgenommen.

Grundsätzlich versucht jedes dieser diskutierten Verfahren die Komplexität der natürlichen Sprache zu reduzieren und somit die Texte für statistische Analysen brauchbarer zu machen, ohne dabei die relevante Informationen zu verringern. Bei Textanalysen gilt der Grundsatz, dass es sehr stark von den spezifischen Textdaten abhängt, welche Vorverabeitungsschritte Sinn machen. Für die Pilotstudie werden deshalb alle möglichen Kombinationen dieser Informationsreduktionsverfahren getestet, was insgesamt 32 verschiedene Textversionen jedes Zeitungsartikels ergibt. Auf diesen Versionen werden die Selektions- und Klassifikationalgorythmen angewandt, welche im nächsten Abschnitt beschrieben werden. Der folgende Auszug zeigt, wie diese Verfahren in R angewendet werden können (das dazugehörige Skript heisst *APS_Klassifikation.r*):

5. Skript-Auszug: Vorverarbeitung

```
# Notwendige Packete laden
library(tm) # Vielseitiges Textberarbeitungstool
library(stringr) # Regular Expressions
library(RTextTools) # Für Textverarbeitung und Textklassifikationen

# Pfad für die Rohtextdateien definieren
directory <- ''raw/''

# Texte in einen Korpus laden (Container für Textkollektionen)
summary(textcorpus <- Corpus(DirSource(directory, encoding = ''UTF-8''), readerControl =
      list(language = ''german'')))

# Dokumentennamen hinzufügen
docnames <- list.files(directory)
for (i in 1:length(textcorpus)){
  meta(textcorpus[[i]], ''Artikel'') <- docnames[i]
}

# Dokument-Term-Matrix kreieren
matrix <- create_matrix(textcorpus, language = ''german'', stripWhitespace = TRUE,
```

```
toLower = TRUE, minWordLength = 3, removeStopwords = TRUE,
removeNumbers = TRUE, stemWords = TRUE, weighting = weigthTf)
```

Zur Vorverarbeitung wird noch ein zusätzliches Paket benötigt, *RTextTools*. Dieses Paket ist speziell für Textklassifikationen entwickelt worden und stellt auch sehr praktische Funktionen für die Textvorverarbeitung zur Verfügung. Nach dem laden von Zeitungsartikeln in Rohtextdateien in einen Textkorpus – in der Linguistik wird eine systematisch geordnete Sammlung von Texten als Korpus bezeichnet –, werden die Artikel in eine Dokument-Term-Matrix transformiert – pro Artikel eine Zeile mit der Vorkommenshäufigkeit der Worte. Wie beschrieben werden alle möglichen Kombinationen der Vorbereitungsoptionen für diese Pilotstudie berücksichtigt, in diesem Skript-Ausschnitt sind beispielhaft forlgende Optionen angewendet: Sprache ist Deutsch, alle Worte in Kleinbuchstaben schreiben, deutsche Stopworte ausschliessen, Zahlen entfernen, und Worte zu Wortstämmen reduzieren. Die auf diese Weise erhaltene Dokument-Term-Matrix kann danach direkt von einer Klassifikationsfunktion eingelesen werden (siehe nächster Abschnitt).

1.1.3 Selektion und Themenklassifikation

Nach der Vorbereitung der Zeitungsartikel erfolgen die eigentlichen Arbeitsschritte, nähmlich die Selektion relevanter Artikel und die Klassifikation der relevanten Artikel in thematische Kategorien. Zusätzlich zu den Optionen der Vorverarbeitung gibt es auf dieser Stufe noch drei weitere Parameter, von denen verschiedene Ausprägungen in einer Pilotstudie getestet werden können. Einen ersten Unterschied kann die Gewichtung der Worte in der Dokument-Term-Matrix (siehe Abschnitt 1.1.2) ausmachen: Entweder es wird auf Worte fokussiert, welche nur in einzelnen Dokumenten häufig sind, oder man fokussiert auf allgemein häufige Worte. Ersteres wird mit einer TF/IDF-Gewichtung[4] erreicht, die eher spezielle Worte stärker gewichtet – mit der Annahme, dass diese relevanter für die Bestimmung der thematischen Kategorien sind. Letzteres bedeutet eine Tf-gewichtung, d.h. dass nach der Gesamtanzahl der Worte in den Dokumenten harmonisiert wird. Diese Unterscheidung wird in *R* implementiert, indem bei der Berechnung der Dokument-Term-Matrix die Option *weighting = weightTfIdf* bzw. *weighting = weightTf* gesetzt wird (vgl. das in Abschnitt 1.1.2 vorgestellte Skript mit Namen *APS_Klassifikation.r*).

Zweitens, in Bezug auf die Klassifizikation von Dokumenten, bietet *RTextTools* neun verschiedene Algorythmen, von welchen drei als effizient gelten. Weil Effizienz für diese Pilotstudie wichtig ist (eine alltagstaugliche Appliaktion müsste in Echtzeit Artikel vorklassifizieren) und die drei effizienten Algorythmen auch meistens gute Leistungen erbringen (v.a. Glmnet), werden die Selektions- und Klassifikationstests mit den folgenden drei Algorythmen durchgeführt (für weitere Informationen über die statistischen Grundlagen der einzelnen Verfahren siehe die Literaturangaben in Klammern):

- SVM: Support vector machine (Meyer, 2012).
- Glmnet: Regularized paths for generalized linear models (Friedman, Hastie and Tibshirani, 2010).
- Maxent: Maximum entropy (Jurka, 2012).

Die verschiedenen Algorythmen können in *R* mit *RTextTools* im Laufe der Klassifikation wie folgt eingesetzt werden (verfügbar im Skript mit Namen *APS_Klassifikation.r*):

6. Skript-Auszug: Klassifikation

```
# Development- und Test-Set definieren
```

[4] TF/IDF = Term Frequency / Inverse Document Frequency

```
container <- create_container(matrix, catH, trainSize = 1:700, testSize = 701:770,
             virgin=FALSE)

# Klassifikationsmodelle am Development-Set trainieren
models <- train_models(container, algorithms=c(''MAXENT'',''GLMNET'', ''SVM''))

# Testset klassifizieren
results <- classify_models(container, models)

# Kennzahlen (Recall, Precision und F-score) extrahieren und anzeigen
analytics <- create_analytics(container, results)
```

Jede Klassifikation (inkl. die Selektion, die als Klassifikation in relevante und irrelevante Artikel verstanden werden kann) läuft prinzipiell nach dem gleichen Schema ab. Zunächst wird die Dokument-Term-Matrix in einen Container geladen, der zwischen den Artikeln des Development-Sets und denjenigen des Test-Sets unterscheidet. Zugleich werden mit der Initialisierung des Containrs die Themen-Codes (hier in der Variable *catH* gespeichert) definiert. An den Artikeln des Development-Sets werden anschliessend die Klassifikationsalgorythmen trainiert. Die trainierten Algorythmen werden dann am Test-Set auf ihre Reliabilität getestet. Die Kennzahlen der Reliabilität (siehe Abschnitt 1.2 für eine Definition und Erklärung) können schliesslich mit dem Befehl *create_analytics* gewonnen werden.

Drittens werden verschiedene Aggregationen des APS-Klassifikationsschemas getestet. Wie in Abbildung 1 dargestellt, nimmt der Schwierigkeitsgrad einer Klassifikation stark zu, wenn die Kategorien feiner werden. Deshalb wird in dieser Pilotstudie der Schweirigkeitsgrad kontinuierlich erhöht, um a) festzustellen, ob eine automatisierte auf der gröbsten Ebene überhaupt gute Resultate leifert und b) inwieweit die automatisierte Klassifikation in Richtung des originalen APS-Kategoriensystems (APS, 2013) verfeinert werden kann. Zu diesem Zweck wurden zwei verschiedene Kategoriensysteme definiert, welche auf den APS-Kategorien basieren, im Vergleich zum Original aber vereinfacht wurden. Dies war notwendig, weil angenommen wird, dass die aktuellen automatisierten Klassifikationsverfahren nicht in der Lage sind, zwischen den feinsten *vierstelligen* Kategorien zu unterscheiden.

Auf der aggregiertesten Ebene der Klassifikation wird versucht, die relevanten Zeitungsartikel in die folgenden Kategorien einzuteilen (Kategoriennummern äquivalent zur originalen Spezifikation (APS, 2013):

Tabelle 2: Gröbere Kategorisierung

10	Staats- und Sozialordnung der Schweiz
11	Institutionen
12	Aussenpolitik
13	Landesverteidigung
14	Finanz- und Geldpolitik
15	Wirtschaftspolitik
16	Sozial- und Bevölkerungspolitik
17	Erziehungs- und Bildungswesen, Medien- und Kulturpolitik
2-8	Gegenstände und Träger der Politik: Kantone und Gemeinden

Die politischen Themen sind hier also auf der obersten Stufe des APS-Kategoriensystems unterschieden. Zudem sind die Kategorien zu den *Trägern* und *Gegenständen* der Politik (Akteure und Institutionen) zu nur einer Kategorie zusammengefasst. Dieses Schema wird als Minimallösung angesehen. Dies bedeutet dass eine Umstellung der APS-Zeitungsdokumentation sich bereits lohnen würde, weil die Fülle an Artikeln nach relevanten gefiltert und grob in Themen zugeteilt werden könnten. Gleichzeitig bedeutet diese Ebene aber auch noch einen erheblichen manuellen Aufwand um die Zeitungsarikel in die endgültigen Kategorien einzuteilen. Deshalb wurde für die Themenbereiche mit den meisten Unterkategorien (Wirtschaftspolitik sowie Sozial- und Bevölkerungspolitik) noch eine ambitioniertere Version des APS-Kategoriensystems erstellt (siehe Tabelle 3), welche grundsätzlich die dreistelligen Codes des APS-Schemas übernimmt, welche aber nach einzelnen Abgrenzungspro-

blemen zusammenfasst wurden.

Tabelle 3: Feinere Kategorisierung

Wirtschaftspolitik	150/151	Allgemeines/Konjunktur
	152	Öffentliche Dienste
	153	Infrastruktur/Energie/Verkehr/Zonenplanung/Umweltschutz
	154/157	Aussenwirtschaft/Handel und Dienstleistungen
	155	Industrie und Gewerbe
	156	Landwirtschaft
Sozial- und Bevölkerungspolitik	160	Bevölkerungsentwicklung
	161	Einwanderungspolitik
	162/163/164	Familienpolitik/Sozialvorsorge/Altersfragen
	165	Sportpolitik
	166	Gesundheitspolitik
	167	Arbeitsrecht

Die zwei in diesem Abschnitt beschreibenen Parameter Gewichtung und Algorythmen wurden in dieser Pilotstudie ebenfalls systematisch getestet. Zu den 32 Versionen aufgrund verschiedener lingusitischer Vorbereitung kommen also für die Tests 6 weitere Variationen hinzu. Insgesamt wurden die Tests also für 192 verschiedene Parameterkombinationen in Bezug auf die Selektion und Klassifikation durchgeführt. Zusätzlich werden die leistungsfähigsten Parameterkombinationen in Bezug die Klassifikation auch noch an den feineren Kategorien getestet.

1.2 Evaluation

Die in den vorherigen Abschnitten beschriebenen Verfahren können für die Periode, für welche die Zeitungsdokumentation des APS digitale Quellen verwendet, getestet werden. Konkret bedeutet dies eine Evaluationsperiode von 1. Januar 2013 bis Mitte November 2013. Für diese Periode sind die Artikel aller vom APS berücksichtigten Zeitungen systematisch und digital nachvollziehbar klassifiziert worden. Das heisst es ist ein Vergleichsdatensatz vorhanden, an welchem die verschiedene Automatisierungen evaluiert werden können. Prinzipiell folgten alle Evaluationen den folgenden Schritten:

- Alle relevanten Zeitungsartikel wurden für alle Zeitungen des APS-Standardsamples heruntergeladen, sofern die Zeitungen im SMD-Archiv vorhanden ist (vgl. Tabelle 10 für eine Auflistung).
- Ziehen einer Zufallsstichprobe von 10'000 Artikeln. Gleichzeitig wurde für die Evaluation der Selektion von Artikeln ein gleich grosses Sample von nicht relevanten Artikeln selektioniert und heruntergeladen. Wie erwähnt verringert sich der Umfang der Samples wegen Downloadproblemen leicht auf 9'921 für die relevanten und 6'688 für die irrelevanten Artikel.
- Der erste Schritt der APS-Zeitungsdokumentation, die Extraktion der Metadaten, ist sehr direkt evaluierbar und benötigt keine besonderen Berechnungen.
- Für die beiden weiteren Schritte der APS-Zeitungsdokumentation (Selektion relevanter Artikel und Klassifikation der Artikel) wurden eingehendere Untersuchungen durchgeführt: Die in Abschnitt 1.1.2 und 1.1.3 beschriebenen Parameter wurden systematisch auf ihre Tauglichkeit überprüft, indem jeweils 90% der Zeitungsartikel in ein Development-Set und 10% der Artikel in ein Test-Set eingeteilt wurden. Das Development-Set dient dazu, die Klassifizierungsverfahren zu trainieren. Am Test-Set wird dann schliesslich die Reliabilität des Verfahrens gemessen.

  $$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

  $$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

  Beide Kennzahlen werden im Vergleich zu einer Referenzcodierung berechnet, in diesem Fall die manuelle Codierung. Im Vergleich zu dieser als richtig angenommenen Codierung sind dann für

alle weiteren Codierungen die *true positives* alle übereinstimmenden Fälle, die *false positives* alle von der Referenzcodierung bestimmten Themenklassifizierungen, die von der Automatisierung nicht gleich bestimmt wurden, und *false negatives* alle Klassifizierungen, die nicht in der Referenzcodierung vorkommen, aber von der automatischen Klassifikation erhoben wurden. Somit ergibt sich ein Bild, wie umfassend (*Recall*) und wie genau (*Precision*) eine Codierung ist. Beide Kennzahlen haben eine potenzielle Spannweite von 0 bis 1, wobei 1 perfekte Reliabilität zwischen der automatisierten und der manuellen Codierung bedeutet.

Aus den beiden Kennzahlen kann schliesslich der F-Score berechnet werden, welcher als harmoniserter Mittelwert sowohl die Ausschöpfungsquote wie auch die Präzision berücksichtigt:

$F = \frac{2 \times Recall \times Precision}{Recall + Precision}$

In der Folge wird die Qualität der Automatisierungen der drei Schritte der APS-Zeitungsdokumentation anhand des *F-Scores* aufgezeigt und diskutiert, wie brauchbar die einzelnen Resultate für eine allfällige Weiterentwicklung der Automatisierungssoftware sind.

1.2.1 Extraktion der Metadaten

Mit den in Abschnitt 1.1.1 beschriebenen Skripte wurden für die Konvertierung der 9'921 relevanten Artikel für die Klassifzierungstests sowie der 9'688 irrelvanten Artikel für die Selektionstests verwendet. Mit einer einzigen Ausnahme, bei der eine Seitennummer nicht erkannt werden konnte, extrahierten die Skripte alle Metadaten (Zeitungstitel, Publikationsdatum, Titel, Seitennummer) einwandfrei.[5] Die gesamten Resulatate können in der Dropbox in den Dateien *meta_classify.txt* sowie *meta_selection.txt* durchgesehen werden. Der F-Score liegt hier deshalb faktisch bei 1 für alle Metadaten. Diese sehr gute Qualität der automatisierten Codierung der Metadaten liegt zu einem grossen Teil an der hohen Konsistenz des SMD-Datenbankformats. Alle Zeitungsartikel aller im SMD indexierten Presseerzeugnisse sind durchwegs einheitlich formatiert, was eine Extraktion der Metadaten einfach gestaltet. Wie in Abschnitt 2.1.1 allerdings beschrieben, ist für diejenigen APS-Quellen, die nicht im SMD vorhanden sind, von einer stärkeren Unstrukturiertheit auszugehen. Mit der Abnahme der Qualität der Ausgangsformate wird aber sicherlich auch die Reliabilität der automatisierten Extraktion der Metadaten etwas leiden. Nichtsdestotrotz ist das Beispiel der Codierung der SMD-Artikel eine starkes Zeichen, dass eine Automatisierung der Extraktion der Metadaten in vollem Umfang alltagstauglich für das APS implementiert werden kann. Eine nachträgliche Überprüfung dieser automatisierten Codierung scheint wenn überhaupt nur noch für einzelne Stichproben notwendig – solange die Konvertierungsskripte bei Nichterkennung Fehlermeldungen produzieren, welche die Inforamtionen für eine schnelle Nachkorrektur liefern.

1.2.2 Selektion der Artikel

Wie gut können Klassifikationsalgorythmen zwischen relevanten und irrelvanten Zeitungsartikeln unterscheiden? Dieser erste Schritt der APS-Zeitungsdokumentation wurde überprüft, indem ein zu den relevanten Artikeln gleich grosses Sample aufgrund einer Zufallsauswahl beschafft wurde (siehe Tabelle 1). Ein Vergleich der vom APS klassifizierten Artikel mit dieser Zufallsauswahl von nicht relevanten Artikeln vermag den Prozess zu simulieren, eine Zeitung manuell nach relevanten Artikeln durchzusehen. Wie erwähnt wurden die in den Abschnitten 1.1.2 und 1.1.3 besprochenen Optionen systematisch getestet. Zunächst folgt deshalb eine Übersicht über die Leistung dieser einzelnen Vorverarbeitungs- und Klassifikationsoptionen. Tabelle 4 zeigt eine Regression der Optionen auf den F-Score der verschiedenen Selektionsläufe.

[5] Die Softwareskripte wurden so ausgelegt, dass sie eindeutige Fehlermeldungen produzieren falls eine Metaangabe nicht erkannt wurde. Dies war aber bis auf die eine erwähnte Seitenangabe nicht der Fall.

Tabelle 4: Einfluss verschiedener Selektionsoptionen auf den F-Score: Koeffizienten, Standardfehler und Signifikanzniveaus von linearen Regressionen

	Deutsch			Französisch		
	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Vorverarbeitungsoptionen						
Stopping	-0.002	0.001	n.s.	0.003	0.003	n.s.
Zahlen	0.001	0.001	n.s.	0.002	0.003	n.s.
Mindestwortlänge	0.000	0.000	n.s.	-0.001	0.001	n.s.
Kleinbuchstaben	0.001	0.001	n.s.	-0.010	0.003	**
Stemming	0.004	0.001	***	-0.005	0.003	n.s.
Gewichtung[a]						
TF/IDF	0.008	0.001	***	-0.019	0.003	***
Algorythmen[b]						
MAXENT	0.008	0.001	***	0.010	0.004	*
SVM	0.014	0.002	***	-0.020	0.004	***
Intercept	0.830	0.002	***	0.726	0.005	***
Fixed effects		Relevanz			Relevanz	
Adj. R^2		0.29			0.20	
N Klassifikationen		320			384	

Anmerkungen:
Signifikanzniveaus: $p \leq 0.001$=***, $p \leq 0.01$=**, $p \leq 0.05$=*.
[a] Referenzkategorie = TF
[b] Referenzkategorie = GLMNET

Im Durchschnitt scheinen die Vorverarbeitungsoptionen für die deutschsprachigen Artikel keine relevanten Unterschiede zu generieren. Im Gegensatz dazu bringen das Stopping und das Stemming für die französischsprachigen Artikel wichtige Verbesserungen. Das Stopping macht ungefähr 5% und das Stemming ungefähr 1,6% des F-Scores aus. In Bezug auf die Gewichtung ist die TF/IDF-Option, welche selten vorkommende Worte stärker gewichtet, wichtig für die Selektion der deutschprachigen Artikel. Die Klassifizierungsalgorhytmen schneiden für die deutsch- und französischsprachigen Artikel gleich ab. Sowohl die *Support Vector Machines* wie auch *Maximum Entropy* sind leistungsfähiger als *Glmnet*. Die im Vergleich besten Resultate erziehlt der *Maximum Entropy* Algorythmus, welcher im Vergleich zu *Glmnet* eine durchschnittliche Leistungssteigerung von 7.5% bzw. 10.9% für die deutsch- bzw. französischsprachigen Artikel ergibt.

Welches ist die maximal erreichte Reliabilität der automatisierten Selektion im Vergleich zur manuellen APS-Selektion und wie stark unterscheiden sich die 192 verschiedenen Klassifikationen? Abbildung 4 zeigt die Streuung des Recalls (d.h. ob alle relevanten Artikel auch selektioniert wurden) und der Precision (d.h. ob noch viele weitere neben den relevanten Artikeln selektioniert wurden) für beide Sprachen. Allgemein machen die verschiedenen Spezifikationen der Selektion keinen so grossen Unterschied aus. Einzig bei der Präzision gibt es in beiden Sprachen einzelne Läufe, welche eine deutlich geringere Leistung als die meisten Klassifikation erreichen. Für die deutschsprachigen Artikel liegt das Maximum sowohl für den Recall wie auch die Precision fast bei 90%, was im Vergleich zu üblichen manuellen Codierungen absolut konkurrenzfähig ist. Eine solcher Wert wird üblicherweise nur für sehr gut geschulte menschliche Codierer erreicht. Für die französischsprachigen Artikel liegt der Wert mit 0.77 knapp unter 0.8, was für immer noch relativ gut ist. Die Unterschiede zwischen den Sprachen könnten unter anderem auch aufgrund von Problemen beim Download der Artikel aus dem

SMD-Archiv entstanden sein. Für die französischsprachigen Artikel wurden einzelne Artikel falsch heruntergeladen.

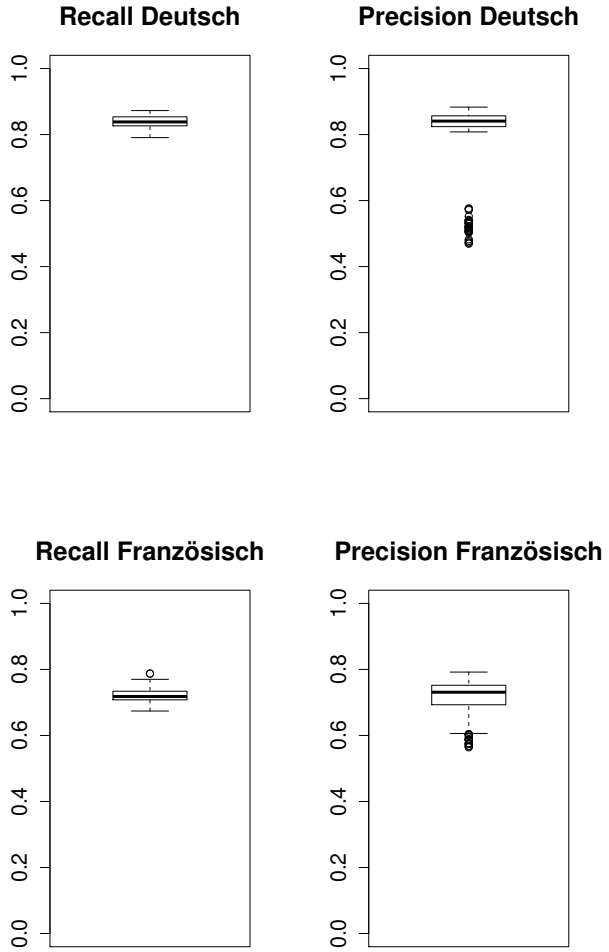

Abbildung 4: Recall und Precision für die Selektion

Die bis hierhin diskutieren relativen Vergleiche sind für eine Pilotstudie sehr relevant, um die Leistung der Klassifizierungen besser untersuchen zu können. Für einen tatsächlichen Einsatz im Alltag einer Zeitungsdokumentation spielen sie aber insofern keine Rolle, weil dafür einfach die beste Kombination gewählt wird. Die folgende Aufstellung gibt Aufschluss darüber, welche Optionen in den Testläufen den besten F-Score erzielten. Der F-Score kann aber in Bezug auf die Selektion nicht unbedingt das relevante Mass sein, weil die Ausschöpfung (= Recall) wahrscheinlich zentraler ist. Genauer gesagt ist es viel wichtiger, dass möglichst alle relevanten Artikel gefunden werden, als dass nicht auch noch einige nicht relevante gefunden würden. Letzteres wird durch einen tiefe Precision ausgedrückt und ist in der Berechnung des F-Scores auch berücksichtigt.

- Französischsprachige Artikel
  - Stopping = Ja
  - Zahlen löschen = Ja
  - Mindestwortlänge = Keine

  - Kleinbuchstaben = Nein
  - Stemming = Nein
  - Gewichtung = TF/IDF
  - Algorythmus = MAXENT
- Deutschsprachige Artikel
  - Stopping = Nein
  - Zahlen löschen = Ja
  - Mindestwortlänge = Keine
  - Kleinbuchstaben = Ja
  - Stemming = Ja
  - Gewichtung = TF/IDF
  - Algorythmus = MAXENT

Wie entwickelt sich die Reliabilität der Selektion, wenn statt über allen Artikeln über einzelene Zeitungen hinweg selektioniert wird? In Tabelle 1 wurden vier Zeitungen identifiziert, für welche mehr als 800 relevante Artikel vorliegen. Für diese Zeitungen wurde die Selektion mit den besten Optionen repliziert, die Resultate sind in Tabelle 7 aufgeführt.

Tabelle 5: Reliabilität verschiedener Selektionen von Zeitungsartikeln

Texte	∅ F-Score
Französisch	
Alle Artikel	0.77
Tribune de Genève	0.76
Deutsch	
Alle Artikel	0.88
Neue Zürcher Zeitung	1.00
Basler Zeitung	0.99
Aargauer Zeitung	0.84

Für die französischsprachigen Artikel ergibt sich keine grosse Verschiebung der Reliabilität. Für die deutschsprachigen Artikel wird für die Basler und Neue Zürcher Zeitung eine fast perfekte Übereinstimmung mit den manuellen Selektionen erreicht. Für die Aargauer Zeitung hingegen resultiert ein leicht schwächerer F-Score. Im Allgemeinen lässt sich aber folgern, dass die automatisierte Selektion zumindest in Bezug auf die deutschsprachigen Artikel gute bis sehr gute Resultate erzielt und in ähnlicher Form beim APS eingesetzt werden könnte, sofern sie durch kontinuierliche Kontrollen einer grösseren Stichprobe ergänzt wird. Für die französischsprachigen Zeitungsartikel müsste beim jetzigen Stand die Selektion nach dem Recall erfolgen, wo einzelne Läufe einem Wert über 0.8 erreichten. Hier sollten die Nachkontrollen, d.h. die Überprüfung der als nicht relevant klassifizierten Artikel, noch intensiviert werden, falls sich die Resultate der automatisierten Selektion nicht verbessern.

1.2.3 Klassifikationen

Der nächste Schritt einer automatisierten E-Dokumentation ist die Unterscheidung der Themen auf der obersten inhaltlichen Ebene. Dazu werden die Parametermöglichkeiten an den in Tabelle 2 gezeigten Kategorien getestet. Wie bei der Selektion werden die Auswertungen zunächst nach Parametern

und dann nach den Themenkategorien untersucht. Anschliessend werden die besten Klassifikationen allgemein sowie einzeln für die Themenkategorien und die wichtigsten Zeitungen diskutiert. Schliesslich folgt der Test am feineren Kategorienschema (siehe Tabelle 3).

Tabelle 6 zeigt die Resultate einer Regression über die verschiedenen Klassifikationsläufe, um den Einfluss der verschiedenenen Klassifikationsoptionen auf den F-Score zu untersuchen. Die Klassifikationen wurden auf der vollen Stichprobe von Artikeln durchgeführt, d.h. auf 7'513 Zeitungsartikel in Deutsch und 2'407 Artikeln in Französisch. Weil nicht alle Klassifikationen erfolgreich waren, ist die Gesamtzahl der Läufe deutlich tiefer als das Maximum von 1'728 (9 inhaltliche Kategorien $\times$ 192 Klassifikationsoptionen). Dies ist nicht unüblich, weil es bei automatisierten Klassifikationen häufig zu Performanzproblemen oder anderen Fehlern kommt, welche die Läufe abbricht. In Bezug auf die Vorverarbeitungsoptionen stechen lediglich die Verbesserungen hervor, welche mit dem Stopping und dem Stemming für die französischsprachigen Artikel erreicht werden können. Für die deutschsprachige Klassifikation sind die Vorverarbeitungsoptionen weniger zentral. Umgekehrtes lässt sich für die Gewichtung sagen, hier ist es für die deutsche Klassifikation wichtig, nach dem TF/IDF-Verfahren zu gewichten, während die Gewichtung im Französischen keinen grossen Unterschied auszumachen scheint. Diese Unterschiede zwischen den Sprachen zeigen deutlich auf, dass es keine allgemein gute automatisierte Textanalyse gibt. Der Erfolg der verschiedenen Verfahren hängt jeweils sehr stark von der Textgrundlage ab. Für die Klassifikationsalgorythmen sind die Resultate schliesslich übereinstimmend: der Maximum-Entropy-Algorythmus ist allgemein für beide Sprachen am geeignetsten. Gegenüber GLMNET erhöht der Einsatz von MAXENT den F-Score durchschnittlich um 7.5% in den deutschsprachigen und um 10% in den französischsprachigen Artikeln.

Tabelle 6: Einfluss verschiedener Klassifikationsoptionen auf den F-Score: Koeffizienten, Standardfehler und Signifikanzniveaus von linearen Regressionen

	Deutsch			Französisch		
	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Vorverarbeitungsoptionen						
Stopping	0.012	0.009	n.s.	0.056	0.011	***
Zahlen	-0.001	0.011	n.s.	-0.002	0.009	n.s.
Mindestwortlänge	0.001	0.003	n.s.	-0.001	0.003	n.s.
Kleinbuchstaben	-0.005	0.008	n.s.	0.004	0.009	n.s.
Stemming	0.013	0.008	n.s.	0.016	0.010	*
Gewichtung[a]						
TF/IDF	0.016	0.007	*	-0.001	0.008	n.s.
Algorythmen[b]						
MAXENT	0.075	0.009	***	0.109	0.010	***
SVM	0.049	0.009	***	0.033	0.010	**
Intercept	0.715	0.014	***	0.482	0.017	***
Fixed effects	Inhaltliche Kategorien			Inhaltliche Kategorien		
Adj. R^2	0.78			0.69		
N Klassifikationen	589			774		

Anmerkungen:
Signifikanzniveaus: p $\leq$ 0.001=***, p $\leq$ 0.01=**, p $\leq$ 0.05=*.
[a] Referenzkategorie = TF
[b] Referenzkategorie = GLMNET

Wie die Abbildung 5 zeigt, gibt es zudem erhebliche Unterschiede in der Klassifikation der einzelnen inhaltlichen Kategorien. In Bezug auf die deutsche Sprache unterscheidet sich die Leistung der Klassifikatoren für die Themen *Staats- und Sozialordnung*, *Institutionen* sowie *Geld- und Finanzpolitik* sehr stark, während die verschiedenen Klassifikationsläufe bei den anderen Kategorien ähnliche Resultate liefern. Für die Precision der deutschsprachigen Klassifikation gibt es grosse Spannweiten bei der Klassifikation der *Geld- und Finanzpolitik* sowie der *Erziehungs-, Bildungs-, Medien- und Kulturpolitik*. Allgemein ist somit zu festzustellen, dass es bei diesen Themen Abgrenzungsprobleme gibt: Die Themen *Staats- und Sozialordnung* und *Institutionen* liegen wohl sehr nahe beieinander, während *Geld- und Finanzpolitik* vermutlich sehr viele Übereinstimmungen mit dem Thema *Wirtschaftspolitik* hat. Das Thema *Erziehungs-, Bildungs-, Medien- und Kulturpolitik* schliesslich, ist wohl in sich zu heterogen und korreliert mit anderen Themen wie z.B. der *Sozialpolitik*. Für das Französische ist das Bild sehr vergleichbar. In Bezug auf die genannten vier inhaltlichen Kategorien bestehen grosse Unterschiede zwischend den Klassifikationen.

In Bezug auf die APS-Zeitungsdokumentation lassen sich aus den Resultaten in Abbildung 5 die folgenden Schlüsse ziehen. Für den Recall, d.h. die Quote mit welcher die in der manuellen Vergleichscodierung den Kategorien zugeordneten Artikel auch richtig zugeordnet wurden, entspricht ein Wert von 0.8 hohen Ansprüchen an die Reliabilität. Dasselbe gilt für die Präzision, d.h. in welcher Zahl die Automatisierte Klassifizierung noch eigentlich nicht relevante Artikel in eine Kategorie eingeteilt hat. Dies kann für den Recell, wenn realistischerweise nur der Maximalwert berücksichtigt wird[6], für sechs inhaltliche Kategorien in Deutsch und für sieben Kategorien in Französisch gesagt werden. In Bezug auf die Präzision sind es drei Kategorien im Deutschen und vier im Französischen. Bei diesen Kategorien ist mit hoher Wahrscheinlichkeit eine manuelle Codierung qualitativ nicht besser. Aber auch die übrigen Kategorien sind mit einer solchen Qualität zugeordnet worden, dass es eine sehr grosse Hilfe bedeuten würde, wenn das APS automatisierte Vorklassifikationen einführen würde. Überhaupt keine Effizienzsteigerung wäre vorhanden, wenn die Klassifikation nicht besser als eine Zufallszuteilung wäre. Eine Zufallszuteilung würde ca. 11% in jede Kategorie einteilen (100% geteilt durch 9 Kategorien). Selbst die schlechtesten Werte sind deutlich höher: Eine F-Score von 0.5 für *Geld- und Finanzpolitik* sowie *Staats- und Sozialordnung* im Französischen sowie ein F-Score von 0.33 für *Institutionen* im Deutschen). Somit ist auf dieser obersten Ebene der inhaltlichen Klassifikation ein erheblicher Effizienzgewinn durch eine Automatisierte Klassifikation zu erwarten, obwohl bei diesem Schritt unbedingt eine Nachkontrolle durch spezialisierte Codierende notwendig ist, um die vereinzelten hohen Ungenauigkeiten zu korrigieren.

[6]In einem realistischen Szenario würden die Zeitungsartikel mit denjenigen Vorbereitungs- und Klassifikationsoptionen durchgeführt, für welche in Tests für die einzelnen inhaltlichen Kategorien die besten Resultate erzielt wurden. Die Vorklassifikation müsste also für jede der 9 inhaltlichen Kategorien alle Artikel klassifizieren.

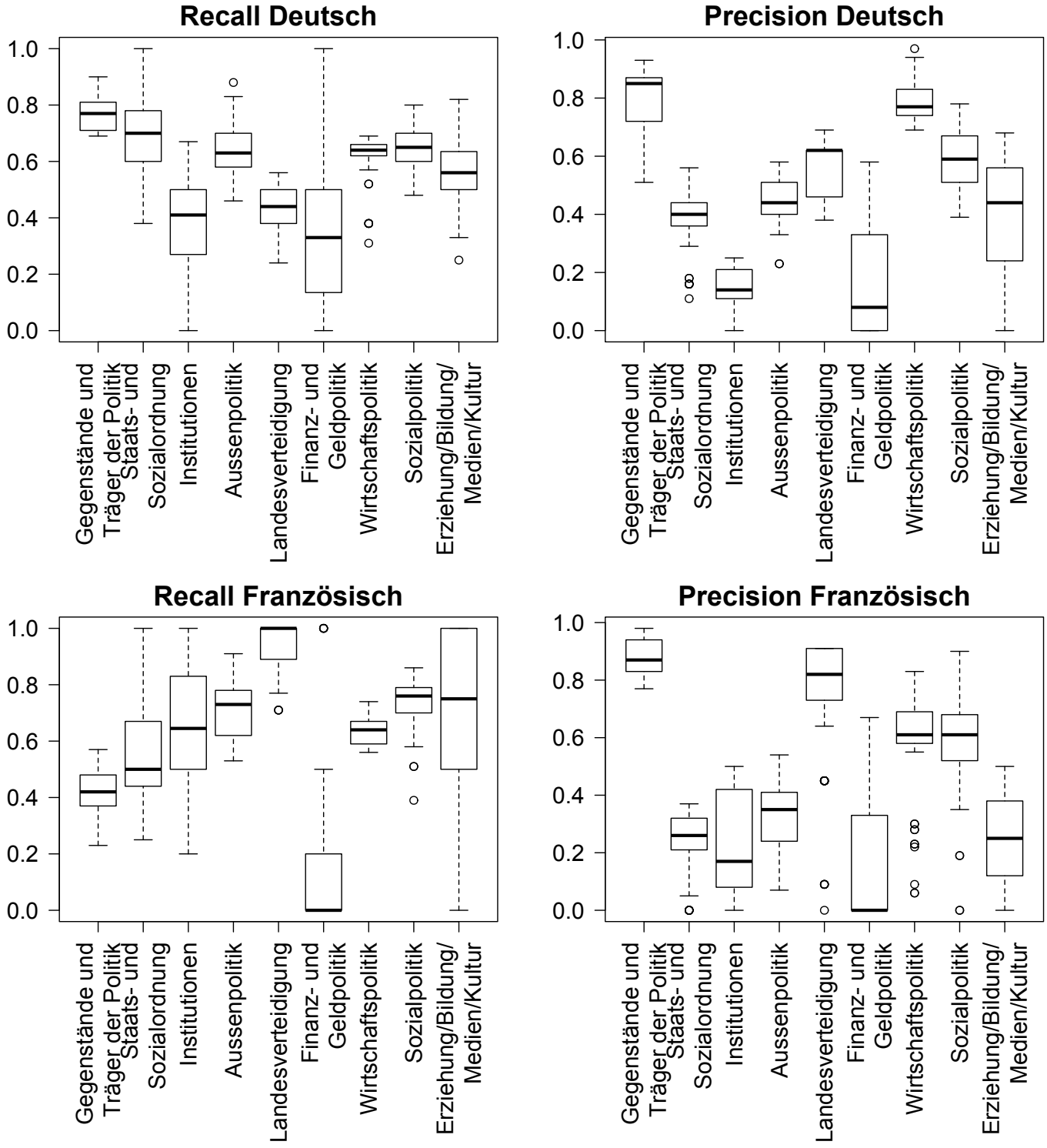

Abbildung 5: Recall und Precision für die inhaltlichen Kategorien

Inwiefern Ändert sich das Bild, wenn entweder auf der Ebende der Zeitungsartikel oder der Ebende der Themenkategorien verfeinert wird? Tabelle 7 zeigt die Veränderung des durchschnittlichen F-Scores für verschiedene Zeitungsartikelselektionen. Als Ausgangspunkt wurde diejenige Klassifikation gewählt, welche im Durchschnitt über alle Themen den höchsten F-Score erzielt. Dies ist die folgenden Kombinationen aus Vorverarbeitungs- und Klassifikationsoptionen:

- Französischsprachige Artikel
  - Stopping = Ja
  - Zahlen löschen = Nein
  - Mindestwortlänge = Keine
  - Kleinbuchstaben = Ja
  - Stemming = Nein
  - Gewichtung = TF
  - Algorythmus = SVM
- Deutschsprachige Artikel
  - Stopping = Ja
  - Zahlen löschen = Nein
  - Mindestwortlänge = 3
  - Kleinbuchstaben = Ja
  - Stemming = Ja
  - Gewichtung = TF/IDF
  - Algorythmus = MAXENT

Wie in Tabelle 1 bereits diskutiert, gibt es vier Zeitungen in der Gesamtselektion, welche mehr als 800 Artikel aufweisen und sich deshalb für einen Test der Klassifikation pro Zeitung besonders eignen.

Tabelle 7: Reliabilität verschiedener Selektionen von Zeitungsartikeln

Texte	∅ F-Score
Französisch	
Alle Artikel	0.64
Tribune de Genève	0.75
Deutsch	
Alle Artikel	0.60
Neue Zürcher Zeitung	0.60
Basler Zeitung	0.58
Aargauer Zeitung	0.62

In Bezug auf die französichen Artikel wird im Vergleich der Klassifikation der Tribune de Genève zu allen französischen Artikeln sogar eine deutliche Steigerung des F-Scores erreicht. Und auch für die deutschsprachige Titel gibt es keine substanzielle Verschlechterung, wenn die Klassifikationen pro Zeitung durchgeführt werden. Klassifikationen könnten deshalb gut auf der Ebene einzelner Zeitungen durchgeführt werden. Tabelle 8 zeigt zudem, inwiefern die Klassifikationen zwischen feineren Klassifikationen unterscheiden können. Als Ausgangsbasis dieses Tests wurden jeweils nur die Artikel der

beiden Kategorien Wirtschaftspolitik und Sozialpolitik (Codes 15 und 16 im APS-Kategorienschema) gewählt. Anschliessend wurde für die oben erwähnte beste Kombination an Klassifikationsoptionen geprüft, inwiefern die in Tabelle 3 aufgelisteten feineren Kategorien automatisiert erkannt werden können. Die F-Scores werden entgegen den Erwartungen deutlich erhöht, für die französischsprachigen um 6%, für die deutschsprachigen Artikel sogar um 16%. Obwohl als Ausgangsbasis dieses Tests nur bereist grobklassifizierte Artikel berücksichtigt wurden, ist dies ein starkes Zeichen, dass die dreistelligen Codes des APS-Kategorienschmemas im Vergleich zu den zweistelligen Kategorien relativ klar voneinander abgegrenzte Klassen beinhalten. Es kann deshalb empfohlen werden, eine allfällige Automatisierung auf der Ebende der dreistelligen Codes durchzuführen.

Tabelle 8: Reliabilität verschiedener Selektionen von Zeitungsartikeln

Texte	∅ F-Score (Anzahl Artikel)
Französisch	
Grobe Kategorisierung	0.60 (2406)
Feinere Kategorisierung	0.76 (667)
Deutsch	
Grobe Kategorisierung	0.64 (7512)
Feinere Kategorisierung	0.70 (2591)

Relative Unterschiede zwischen den Vorbereitungsoptionen und Themen sind wichtig, um die Stärken und Schwächen der Klassifikation zu bestimmen. Für einen tatsächlichen Einsatz einer automatisierten Vorklassifikation würden schliesslich aber nur diejenigen Klassifizeirungsläufe zum Zuge kommen, welche pro Thema die besten Resultate erzielen. Tabelle zeigt diese besten Klassifikationen anhand des F-Scores pro Thema. Bei einer Themenklassifikation ist der F-Score ein sinnvolles Mass, weil er ein Gleichgewicht zwischen nicht zugeteilten und falsch zugeteilten Artikeln anzeigt und somit generell die Fehler minimieren hilft. Die Labels der Themenkategorien sind in Tabelle 2 ablesbar. Wie bereits erwähnt zeigen diese Resultate, dass es keine allgemeingültig beste Klassifikation gibt, sondern dass die Optionen der Klassifikation bei jeder neuen Aufgabe wieder neu eruiert werden müssen.

Tabelle 9: Anhand des F-Scores bestimmte beste Klassifikationen pro Thema

Thema	F-Score	Stop-ping	Zahlen	Mindest-wortlänge	Kleinbuch staben	Stem-ming	Gewichtung	Algo-rythmus
Deutschsprachige Artikel								
10	0.63	Ja	Nein	Nein	Nein	Ja	TF	SVM
11	0.33	Ja	Ja	Nein	Ja	Ja	TFIDF	MAXENT
12	0.64	Ja	Ja	Nein	Ja	Ja	TFIDF	MAXENT
13	0.57	Nein	Ja	Nein	Ja	Nein	TF	SVM
14	0.57	Nein	Ja	Nein	Ja	Nein	TFIDF	MAXENT
15	0.76	Ja	Ja	Nein	Ja	Ja	TFIDF	MAXENT
16	0.69	Nein	Nein	3	Ja	Nein	TFIDF	MAXENT
17	0.68	Nein	Nein	Nein	Nein	Ja	TF	SVM
2	0.84	Nein	Nein	Nein	Ja	Nein	TF	SVM
Französischsprachige Artikel								
10	0.50	Ja	Nein	Nein	Ja	Nein	TF	SVM
11	0.60	Ja	Nein	Nein	Ja	Nein	TF	SVM
12	0.67	Nein	Ja	Nein	Ja	Ja	TFIDF	MAXENT
13	0.95	Ja	Ja	Nein	Nein	Nein	TF	SVM
14	0.50	Ja	Ja	Nein	Nein	Nein	TFIDF	SVM
15	0.75	Nein	Nein	Nein	Nein	Nein	TFIDF	MAXENT
16	0.76	Ja	Nein	Nein	Ja	Nein	TFIDF	SVM
17	0.62	Ja	Ja	Nein	Nein	Nein	TF	SVM
2	0.68	Nein	Nein	3	Ja	Nein	TFIDF	MAXENT

2 Empfehlungen für eine Neuorganisation der Zeitungsdokumentation

2.1 Prozessabläufe

Die Pilotstudie hat ergeben, dass automatisierte Verfahren eine beträchtliche Hilfe für die APS-Zeitungsdokumentation darstellen könnten. Besonders für die Erkennung der Metadaten und die erste Auswahl an relevanten Zeitungsartikeln bergen automatisierte Verfahren das Potenzial, die Effizienz bei mindestens gleichbleibender Qualität erheblich zu steigern. Mit der gesteigerten Effizienz könnte das APS Kapazitäten für andere Arbeiten schaffen, mehr Quellen in die Zeitungsdokumentation aufnehmen oder die Qualitätskontrolle intensivieren (siehe Abschnitt 2.2). Damit dieses Potenzial ausgeschöpft werden kann, sind aber Änderungen an den Arbeitsprozesse notwendig.

Die vielen Arbeitschritte der APS-Zeitungsdokumentation lassen sich grundsätzlich in drei Abschnitte unterteilen: 1) Beschaffung der Daten und Erhebung der Metadaten, 2) automatisierte Vorbereitung und vorläufige Klassifikation der Zeitungsartikel und 3) endgültige Klassifikation und Sicherung der Daten. Der ganze Prozess wird am besten als halbautomatische Codierung verstanden und durch eine relationale Datenbank (SQL) sowie einem Codier-Interface für die Erhebung der Metadaten sowie die endgültige Kontrolle unterstützt (siehe grau hinterlegte Elemente in Abbildung 6). Eine solche Softwareinfrastruktur wird anhand eines extra für diesen Bericht erstellten Prototyps aufgezeigt, welcher als mögliches Vorbild für eine später einsatzfähige Codierapplikation dienen kann.

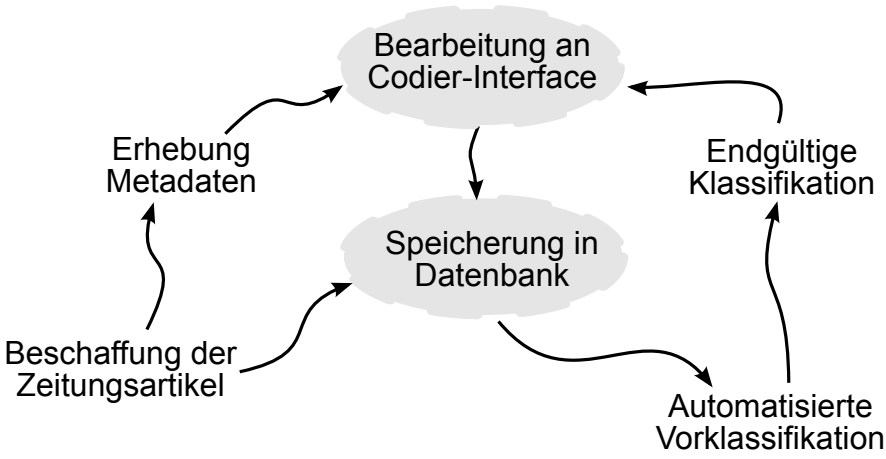

Abbildung 6: Grunsätzliche Prozessabläufe einer halbautomatischen Zeitungsdokumentation

2.1.1 Beschaffung der Artikel und Erhebung der Metadaten

Die Beschaffung von Artikeln wird sich stark nach den Lizenzbedingungen der vom APS benötigten Presseerzeugnissen richten. Aber es wird selbst bei geringen Umstellungen der Beschaffungsarten möglich sein, die Metadaten zumindest teiweise automatisiert zu codieren. Grundsätzlich gibt es zwei Möglichkeiten, wie die Zeitungsartikel kontinuierlich in digitaler Form zum APS gelangen können: über direkte Abonnemente oder über Datenbanken. Erstens können wie bisher Abonnemente direkt bei den Zeitungen bestellt werden. Dies hat allerdings zwei gewichtige Nachteile: Erstens liefert jede Zeitung die Artikel in einem anderen Format (z.B. verschiedene pdf-Formate, unterschiedliche XML- und HTML-Definitionen etc.). Das bedeutet, dass für jede Quelle wieder ein neues Konvertierungsprogramm, welches die eintreffenden Artikel wie in Abschnitt 1.1.1 beschrieben in ein Standardformat umwandelt, geschrieben werden muss. Zweitens liefern viele Zeitungen die Artikel im pdf-Format. Pdf-Dateien sind entweder eingebettete Bilder, die überhaupt nicht digital weiterverarbeitet werden können (z.B. liefert *Der Bund* seine Artikel in dieser Form), oder aber über Distanzen und Abstände gesetzter Text. Letzteres Format kann zwar in eine Rohtextdatei konvertiert werden, wegen dem speziellen Setzprinzip bei pdf-Formaten wird es aber immer Probleme geben. Dies kann anhand der Titelseite der Basler Zeitung vom 10.12.2013 veranschaulicht werden (siehe Abbildung 7). Diese Titelseite wurde mit einem Standardprogramm (siehe `http://www.extractpdf.com/`) in Text konvertiert und müsste für den weiteren Gebrauch noch stark manuell nachbearbeitet werden, u.a. weil der Zeitungstitel irgendwo im Impressum aufgeführt ist, die einzelnen Schlagzeilen geschachtelt im Hauptartikel enthalten sind und die Konvertierung falsche Leerzeichen im Wort *Fiko* eingefügt hat.

Basler Zeitung

Dienstag, 10. Dezember 2013 | Fr. 3.–
(inkl. MWSt)
Nummer 296 | 171. Jahrgang
Basler Zeitung | Aeschenplatz 7 | Postfach 459 | 4010 Basel
Tel. 061 639 11 11 | Fax 061 631 15 82 | E-Mail redaktion@baz.ch
Abonnements- und Zustelldienst: Hochbergerstrasse 15, Postfach, 4002 Basel,
Tel. 061 639 13 13 | E-Mail abo@baz.ch
Elsass/Deutschland € 2.80

Meinungen/Profile/Impressum 8–9 Region 11–17 Notfälle 18 Bestattungen 18 Wetter 20 Kultur 21–30 Fernsehen/Radio 28–29 Wirtschaft 31–33 Kino 32 Börse 34–35 Sport 36–40

Schweiz

Steuerstreit. Der Bundesrat plädiert für mehr Vertrauen in die USA und setzt auf ein baldiges Ende. Seite 4

«Gefährlich». Die Regierung lehnt die Volksinitiative «Abtreibungsfinanzierung ist Privatsache» vehement ab. Seite 5

International

Jungstar. Der neue Chef von Italiens regierender Demokratischer Partei (PD), Matteo Renzi, ist für viele ein Hoffnungsträger – auch bei Anhängern der Rechten. Seite 6

Basel

Rheinufer. Das Komitee «Unser Stadtbild» will den geplanten Steg am Grossbasler Ufer verhindern. Seite 11

Apfelklau. Ein Oberdörfer Apfelbaumbesitzer bemüht wegen gestohlener Früchte die Justiz. Seite 17

BVB-Präsident Gudenrath muss gehen

Regierungsrat Wessels entzieht dem obersten Drämmler nach Filzvorwürfen das Vertrauen

Von Aaron Agnolazza und Daniel Wahl

Basel. Martin Gudenrath, Verwaltungsratspräsident der Basler Verkehrsbetriebe (BVB), nimmt den Hut. Nachdem der Bau- und Verkehrsdirektor Hans-Peter Wessels dem BVB-Präsidenten noch im September sein volles Vertrauen ausgesprochen hat, entzieht er ihm dieses jetzt. Grund für die Kehrtwende Wessels ist der Bericht der Finanzkontrolle, der am Samstag an alle Verwaltungsratsmitglieder verschickt und in der gestrigen Sitzung behandelt wurde.

Die im Raum stehenden Vorwürfe an die Adresse von Gudenrath und mehrere Mitglieder der BVB-Leitung, darunter auch Direktor Jürg Baumgartner, waren gravierend: Wie die BaZ im August publik machte, ging es dabei um Vetternwirtschaft und Begünstigung bei der Leitung der BVB. So seien für den Sohn von Baumgartner sowie die Kinder von Vizedirektor Franz Brunner und VR-Präsident Martin Gudenrath Praktikumsstellen neu geschaffen worden, für welche die Sprösslinge überdurchschnittlich entlöhnt worden sind.

Weitere Vorwürfe betrafen Regelungen zur privaten Nutzung von Dienstwagen, Antrittsprämien sowie Unterkunftsspesen für eine Wohnung des im zürcherischen Ottenbach wohnhaften Baumgartner. In der Folge wurde die Finanzkontrolle (Fiko) mit einer unabhängigen Untersuchung beauftragt und auch die Geschäftsprüfungskommission des Grossen Rats schloss eine Überprüfung der BVB nicht aus.

Paul Rüst, Vizepräsident des BVB-Verwaltungsrats, der die gestrige Sitzung bereits leitete, bestätigt die Vorwürfe, ohne auf die Inhalte des Berichts der Fiko konkret einzugehen. Es seien Kompetenzen überschritten worden und rechtliche Grundlagen ohne die gesetzlich vorgesehene Zustimmung der Personalkommission erlassen worden. Auch die Arbeitsweise des gesamten Verwaltungsrats müsse ab Januar auf neue Beine gestellt werden, um eine verbesserte Kontrollfunktion wahrnehmen zu können, sagt Rüst. Der Fokus werde auf die «Rechtsgrundlagen» gerichtet, die für mehr Klarheit sorgen und Differenzen in der Betrachtungsweise von Jürg Baumgartner ausräumen sollten. Zur Überwachung der Massnahmen, die die Fiko vorgeschlagen hatte, setzt der Verwaltungsrat einen Ad-hoc-Ausschuss ein. Mit Martin Gudenrath geht auch Dominik Egli als Präsident des Verwaltungsratsausschusses Finance & Compliance. Heute Dienstag wird Wessels dem Gesamtregierungsrat, der drei Mitglieder des Verwaltungsrats wählen muss, nun auch einen neuen Präsidenten vorschlagen.

```
Dienstag, 10. Dezember 2013 | Fr. 3.–

(inkl. MWSt) Nummer 296 | 171. Jahrgang Basler Zeitung | Aeschenplatz 7 | Postfach 459 | 4010 Basel Tel. 061 639 11 11 | Fax 061 631 15 8
Postfach, 4002 Basel, Tel. 061 639 13 13 | E-Mail abo@baz.ch Elsass/Deutschland € 2.80

Meinungen/Profile/Impressum 8–9  Region 11–17  Notfälle 18  Bestattungen 18  Wetter 20  Kultur 21–30  Fernsehen/Radio 28–29  Wirtschaft 3

Schweiz
Steuerstreit. Der Bundesrat plädiert für mehr Vertrauen in die USA und setzt auf ein baldiges Ende.  Seite 4 «Gefährlich». Die Regierung
Seite 5

BVB-Präsident Gudenrath muss gehen
Regierungsrat Wessels entzieht dem obersten Drämmler nach Filzvorwürfen das Vertrauen
Von Aaron Agnolazza und Daniel Wahl
Basel. Martin Gudenrath, Verwaltungs-

International
Jungstar. Der neue Chef von Italiens regierender Demokratischer Partei (PD), Matteo Renzi, ist für viele ein Hoffnungsträger – auch bei A

Basel
Rheinufer. Das Komitee «Unser Stadtbild» will den geplanten Steg am Grossbasler Ufer verhindern.  Seite 11 Apfelklau. Ein Oberdörfer Apfe
Die Chancen, dass der Landrat an der Gebärstation im Spital Laufen festhält, sind gering.  Seite 17

ratspräsident der Basler Verkehrsbetriebe (BVB), nimmt den Hut. Nachdem der Bau- und Verkehrsdirektor Hans-Peter Wessels dem BVB-Präsiden
dieses jetzt. Grund für die Kehrtwende Wessels ist der Bericht der Finanzkontrolle, der am Samstag an alle Verwaltungsratsmitglieder vers
Vorwürfe an die Adresse von Gudenrath und mehrere Mitglieder der BVB-Leitung,

darunter auch Direktor Jürg Baumgartner, waren gravierend: Wie die BaZ im August publik machte, ging es dabei um Vetternwirtschaft und Be
die Kinder von Vizedirektor Franz Brunner und VR-Präsident Martin Gudenrath Praktikumsstellen neu geschaffen worden, für welche die Sprös
Regelungen zur privaten Nutzung von Dienstwagen, Antrittsprämien sowie Unterkunftsspesen für eine Wohnung des im zürcherischen Ottenbach

F  inanzkontrolle (Fiko) mit einer unabhängigen Untersuchung beauftragt und auch die Geschäftsprüfungskommission des Grossen Rats schloss
BVBVerwaltungsrats, der die gestrige Sitzung bereits leitete, bestätigt die Vorwürfe, ohne auf die Inhalte des Berichts der F  iko konkre
ohne die gesetzlich vorgesehene Zustimmung der Personalkommission erlassen worden. Auch die Arbeitsweise des gesamten Verwaltungsrats müs

besserte Kontrollfunktion wahrnehmen zu können, sagt Rüst. Der Fokus werde auf die «Rechtsgrundlagen» gerichtet, die für mehr Klarheit so
sollten. Zur Überwachung der Massnahmen, die die Fiko vorgeschlagen hatte, setzt der Verwaltungsrat einen Ad-hoc-Ausschuss ein. Mit Marti
Finance & Compliance. Heute Dienstag wird Wessels dem Gesamtregierungsrat, der drei Mitglieder des Verwaltungsrats wählen muss, nun auch
```

Abbildung 7: Probleme bei der Transformation von pdf- in Textdateien (oben = pdf-Original; unten = konvertierte txt-Datei)

Solche Fehler in der Anordnung der Textbestandteile einer pdf-Seite sind keineswegs Einzelfälle und machen eine Anwendung von pdf-Dateien für eine digitale Bearbeitung sehr problematisch. Konkret scheitert eine Automatisierung bei solch beschädigten Texten bereits am grossen Aufwand für die Textvorbeitung. Wenn Abonnemente bei einzelnen Zeitungen bezogen werden, sollte deshalb unbedingt abgeklärt werden, ob diese Zeitungen auch Artikel in einem besseren Format (Text, XML, HTML oder Office-Dateien) geliefert werden können. Die Zeitungen haben intern digital besser lesbare Textformate. Ob Sie diese veröffentlichen ist aber unklar und müsste vom APS aktiv ausgehandelt werden.

Eine zweite Möglichkeit ist, die Zeitungsartikel über Datenbanken zu beschaffen. Dies aus zwei Gründen eindeutig zu bevorzugen. Erstens sind die Texte in allen bekannten Datenbanken in einem digital gut bearbeitbaren Format vorhanden (XML oder HTML). Zweitens speichern die Datenban-

ken die Zeitungsartikel systematisch im gleichen Format ab. Es braucht also nur noch ein Konvertierungsprogramm pro Datenbank. Die wichtigste Datenbank für Schweizer Presseerzeugnisse, die auch in der Pilot-Studie (siehe Abschnitt 1) verwendet wurde, ist der Schweizerische Mediendienst (SMD, `www.smd.ch`). Tabelle 10 zeigt die Übereinstimmungen zwischen dem Datenbestand des SMD und der APS-Auswahl von Standardzeitungen.

Tabelle 10: Verfügbarkeit von Zeitungen in der SMD-Datenbank

24 Heures	Ja	Neue Urner Zeitung*	Nein
Aargauer Zeitung	Ja	Neue Zuger Zeitung*	Nein
Appenzeller Zeitung*	Nein	Neue Zürcher Zeitung	Ja
Basellandschaftliche Zeitung	Ja	Nouvelliste	Ja
Basler Zeitung	Ja	Quotidien Jurassien	Nein
Berner Zeitung	Ja	Schaffhauser Nachrichten	Ja
Blick	Ja	Solothurner Zeitung	Ja
Bote der Urschweiz	Nein	Sonntags-Zeitung	Ja
Bund	Ja	Sonntagsblick	Ja
Der Sonntag	Nein	St. Galler Tagblatt	Ja
L'Express	Ja	Südostschweiz	Ja
La Liberté	Ja	Südostschweiz Glarus*	Nein
Le Matin	Ja	Tages-Anzeiger	Ja
Le Temps	Ja	Thurgauer Zeitung	Ja
Neue Luzerner Zeitung	Ja	Walliser Bote	Ja
Neue Nidwaldner Zeitung*	Nein	Weltwoche	Ja
Neue Obwaldner Zeitung*	Nein	Wochenzeitung	Ja

* Regionalausgaben

Die Aufstellung in Tabelle 10 zeigt, dass die überwiegende Mehrheit der vom APS am meisten benötigten Titel (28 von 37) im SMD indexiert sind. Zudem ist die Möglichkeit gross, dass die sechs Regionalausgaben (mit $*$ gekennzeichnet) der bereits vorhandenen Zeitungen (Neue Luzerner Zeitung für Neue Zuger, Urner, Obwaldner und Nidwaldner Zeitung; Südostschweiz bei Südostschweiz Glarus bzw. Thurgauer Zeitung für die Appenzeller Zeitung) ebenfalls in der SMD-Datenbank indexiert sind. Dies wurde im Zuge dieses Berichts nicht abschliessend geklärt und Bedarf deshalb der Kontaktaufnahme mit dem SMD. Der SMD ist ein Joint Venture zwischen der SRG, Ringer und Tamedia, ein potenzieller Zugang des APS unterliegt deshalb der Zustimmung dieser drei Eigentümer, auch inwiefern eine automatisierte Beschaffung möglich ist (z.B. über einen API-Zugang[7]). Es gibt weitere Datenbanken, welche einige der Lücken beim SMD schliessen könnten. Die Universität Bern verfügt über einen Bibliothekszugang zu den Datenbanken Lexis-Nexis[8] und Factiva[9], welche einige Schweizer Presseerzeugnisse indexiert hat. Dies wäre eine gute Verhandlungsgrundlage, um mit den genannten Anbietern einen API-Zugang zu verhandeln. Allerdings sind solche Zugänge im Vergleich zum SMD wahrscheinlich teurer. Ein API-Zugang bei Lexis-Nexis zum Beispiel beläuft sich nach den aktuellsten verfügbaren Angaben auf CHFr. 9000.- im ersten Jahr und dann CHFr. 3000.- für jedes weitere Jahr.

Wie in Abschnitt 1.1.1 am Besipiel von Artikeln aus der SMD-Datenbank beschrieben, müssen die Artikel nach der Beschaffung zunächst standardisiert werden. Dieser Schritt beinhaltet die Extraktion des Rohtextes aus den Originaldokumenten, die Codierung der Metadaten (Zeitungsname, Erscheinungsdatum und Titel des Artikels) sowie die Einspeisung der Artikel in eine Datenbank.

2.1.2 Verantwortlichkeiten und Kategorienschemata

Wenn Atomatisierungen für die APS-Zeitungsdokumentation ins Auge gefasst werden, hat dies erhebliche Auswirkungen auf die Arbeitsprozesse. Grundsätzlich wird sich die hauptsächliche Arbeits-

[7] API = Advanced Programming Interface, eine Softwarelösung, die einen programmierbaren Anschluss an die Datenbank und somit einen automatisierten Download aller Zeitungsartikel ermöglicht.

[8] `http://e-solution.lexisnexis.de/KSH/en/index.html`

[9] `http://www.ub.unibe.ch/content/suchen__finden/datenbanken/index_ger.html?id=964`

last vom eigentlichen Erfassen der Zeitungsartikel zur Kontrolle der automatisierten Klassifikation verschieben. Eine automatisierte Klassifikation wird immer Fehler produzieren. Deshalb wird die Nachbearbeitung der Vorklassifikation ein wesentlicher Bestandteil der APS-Zeitungsdokumentation bleiben, ungeachtet der Tatsache wie viel automatisiert wird. Eine konkrete Änderung, welche bei einem Einsatz automatisierter Vorstufen empfehlenswert ist, wäre die Aufteilung der Kompetenzen nicht wie bisher nach Zeitungen, sondern nach den APS-Themenkategorien. Die automatisierte Vorcodierung gibt bereits vorklassifizierte Artikel aus, weshalb es sinnvoll ist, wenn es ThemenspezialistInnen die vorklassifizierten Artikel kontrollieren und allenfalls umteilen. Bei einer solchen Änderung der Abläufe sollte darauf geachtet werden, dass die zugeteilten Themenblöcke nicht zu heterogen sind. Aus den Evaluationen ging beispielsweise hervor, dass die Themen *Staats- und Sozialordnung* und *Institutionen* sowie *Geld- und Finanzpolitik* sowie *Wirtschaftspolitik* wohl nicht gut voneinander abzugranzen sind durch einen automatisierte Vorcodierung. Diese Themen sollten deshalb dem gleichen Codierenden zugetailt werden.

2.1.3 Zusammenspiel zwischen automatisierter und manueller Codierung

Wie bereits mehrfach erwähnt werden auch längerfristig bei einer Automatisierung manuelle Eingriffe durch ExpertInnen der APS-Zeitungsdokumentation notwendig bleiben. Einerseits muss die automatisierte Vorklassifikation kontrolliert und allenfalls korrigiert und ergänzt werden, weil nur intellektuelle Entscheide der Codierenden mit den feinsten Themenkategorien in einer zufriedenstellenden Qualität umgehen können. Andererseit brauchen die Schätzverfahren der automatisierten Vorklassifikation laufend aktualisierte Vergleichsdaten. Letzteres bedeutet, dass die Codierapplikation die endgültige manuelle Klassifikation fortlaufend als Grundlage für die Vorklassifikation der neuen Artikel benutzen sollte. Dies ist nicht nur zentral, weil mehr manuell kontrollierte Artikel aufgrund der grösseren Datenmengen eine bessere Qualität der Vorklassifikation bedeutet, sondern auch weil sich die Sprache in den Zeitungsartikeln über die Zeit verändert. Durch eine konstante Aktualisierung der Vergleichsdaten kann die automatisierte Vorklassifikation mit diesen Veränderungen Schritt halten. In letzter Konsequenz bedeutet dies, dass die Parameter der Vorklassifikation laufend angepasst werden sollten, um ein optimales Gleichgewicht zwischen Präzision und Ausschöpfung erreicht werden kann. Die automatisierten Klassifikationen – vor allem wenn sie mit einem Test aller möglichen Klassifizierungsoptionen einhergehen – können allerdings sehr lange Zeit beanspruchen. Bespielsweise dauerte die Klassifikation aller deutschen Artikel ungefähr 51 Stunden. Aus diesem Grund wird empfohlen, die Klassifikatione auf der Ebene der einzelnen Zeitungen zu aktualisieren und nur sporadisch das ganze Sample zu testen.

Diese Prozesse können mit einer Codierapplikation gesteuert werden. Im folgenden wird ein Prototyp einer mit *shiny* und der MySQL-Anbindung in *R* (*RMySQL*) aufgebauten Applikation vorgestellt, welche ganz konformtabel über einen Browser angesteuert werden kann. *Shiny* wird für den Aufbau der Serverappliaktion benötigt, *RMySQL* wird für die Bedienung einer relationalen Datenbank benötigt, welche die Metadaten, alle Textversionen und die Klassifikationen der Artikel enthält. Dieser Prototyp ist für einzelne Teilaufgaben funktionstüchtig und soll eine Möglichkeit der Umsetzung aufzeigen. Er ist aber noch in keinster Weise einsatzfähig, weil er nicht an grösseren Mengen von Zeitungsartikeln getestet worden ist, die Eingaben am Interface noch nicht aufgefangen werden, und das Interface weder mit der in der Pilotstudie verwendeten Klassifikationssoftware noch der SQL-Datenbank verbunden wurde.

Zunächst wird eine Datenbank mit den in Tabelle 11 aufgelisteten Tabellen und Variablen definiert. Es werden vier Tabellen definiert, wobei die ersten drei Tabellen die Kontextinformationen zu den Zeitungen, Codierenden sowie Themen enthalten. Die vierte Tabelle enthält die Zeitungsartikel und referenziert über die jeweiligen Identifikationsnummern auf die ersten drei Tabellen. Somit können zu jedem Artikel die entsprechenden Kontextinformationen aus den Tabellen geholt werden.

Tabelle 11: Definition Datenbank

Variablenname	SQL-Datentyp	Beschreibung
Tabelle Quellen		
source_ID	integer	Identifikationsnummer des Zeitungstitels
source_name	varchar	Name der Zeitung
source_init	timestamp	Datum und Zeit der Erstellung des Eintrags
APS_shortname	varchar	Vom APS verwendetes Kürzel
SMD_shortname	varchar	Vom SMD verwendetes Kürzel
Tabelle Coder		
coder_ID	integer	Identifikationsnummer des Codierenden
coder_name	varchar	Name des Codierenden
coder_init	timestamp	Datum und Zeit der Erstellung des Eintrags
Tabelle Themen		
issue_ID	integer	Identifikationsnummer des Themas
issue_name	varchar	Name des Themas
APS_code	varchar	Im APS verwendete Identifikationsnummer
issue_init	timestamp	Datum und Zeit der Erstellung des Eintrags
issue_descr	text	Beschreibung des Themas
Tabelle Artikel		
art_ID	bigint	Identifikationsnummer des Artikels
source_ID	integer	Referenz auf Identifikationsnummer des Zeitungstitels
coder_ID	integer	Referenz auf Identifikationsnummer des Codierenden
issue_ID_final	integer	Finale Einteilung des Artikels; Referenz auf ID des Themas
issue_ID_auto	integer	Vorschlag der automatisierten Vorcodierung; Referenz auf ID des Themas
art_init	timestamp	Datum und Zeit des Rinlesens des Artikels
pub_date	date	Publikationsdatum des Artikels
title	varchar	Titel des Artikels
page	varchar	Seitennummer des Artikels
text	text	Text des Artikels

In *R* kann eine solche Datenbank wie im folgenden Auschnitt dargestellt definiert werden (vgl. das Skript *APS_Datenbank.r* für den gesamte Software-Code):

7. Skript-Auszug: Definition Datenbank

```
# Datenbank-Verbindung aufbauen (Eine Datenbank mit Namen 'APS' wurde bereits definiert)
con <- dbConnect(MySQL(), dbname='APS-Repo', user='<Benutzername>', pass='<Passwort>', host=
        'localhost', port=8889)

# Zeitungstabelle definieren
dbSendQuery(con, ''create table sources (source_ID int not null auto_increment primary key,
  source_name varchar(25) not null, source_init timestamp not null, APS_shortname varchar(25)
  not null, SMD_shortname varchar(25) not null, unique (source_name))'')

# Codertabelle definieren
dbSendQuery(con, ''create table coders (coder_ID gint not null auto_increment primary key,
  coder_name varchar(25) not null, coder_init timestamp not null, unique (code_name)'')

# Thementabelle definieren
dbSendQuery(con, ''create table issues (issue_ID int not null auto_increment primary key,
  issue_name varchar(25) not null, APS_code varchar(25) not null, issue_init timestamp not
  null, issue_descr text not null, unique (issue_name))'')

# Artikeltabelle definieren
dbSendQuery(con, ''create table articles (art_ID bigint not null auto_increment primary
  key, source_ID int not null, coder_ID int not null, issue_ID_final int not null,
  issue_ID_auto int not null, art_init timestamp not null, pub_date date not null, title
  varchar(250) not null, page varchar(25) not null, text text not null, unique (art_ID))'')
```

Zunächst wird eine Verbindung zur MySQL-Datenbank mit dem Namen *APS-Repo* aufgebaut (*con*).

Die Option *localhost* kann gesetzt werden, wenn die Datenbank auf dem gleichen Computer wie das Skript liegt, welches ausgeführt wird. Wenn nicht, muss die IP des Servers, wo die SQL-Datenbank abgelegt ist, angeben werden. Dann werden viermal SQL-Befehle an die Datenbank gesendet, die alle die genannten Tabellen definieren, mitsamt den in Tabelle 11 aufgeführten Variabalen und deren Spezifikation (z.B. Variablentyp, automatische Vervollständigung (*auto_increment*) oder Eindutigkeit (*unique(...)*).

Nach dem kreieren der Datenbanken könne diese mit Inhalt versehen werden. Der nachfolgende Ausschnitt zeigt diesen Schritt an einem einfachen Beispiel (siehe das Skript *APS_Datenbank.r*):

8. Skript-Auszug: Datenbank-Update I

```
# Zeitungstabelle aktualisieren
sources <- read.table(''sources.txt'', header = T, quote = '', fileEncoding = ''UTF-8'',
          sep = ''\t'')
sources <- subset(sources, select = c(''Name'', ''AbbrAPS'', ''AbbrSMD'', ''Coder''))
colnames(sources) <- c(''source_name'', ''APS_shortname'', ''SMD_shortname'',
                       ''coder_ID'')
dbWriteTable(con, ''sources'', sources, append = T, row.name = F)
```

Die Tabelle, welche die Angaben zu den Zeitungen (Verwendete Kurznamen, zugewiesene Codierende etc.) enthält, kann zum Beispiel in einem Spreadsheet definiert werden (hier mit Namen *sources.txt*). Nach dem einlesen in *R* und der Anpassung der variablennamen an die Datenbank-Definitionen kann die Tabelle einfach in SQL eingelesen werden. Für die Artikel muss eine komplexere Aktualisierung definiert werden, wie im folgenden Skript-Auszug ersichtlich ist.

9. Skript-Auszug: Datenbank-Update II

```
# Artikeltabelle aktualisieren
# Liste der Dateinamen der konvertierten Artikel definieren
directory <- ''2_read_test/''
docnames <- list.files(directory)

# Einen Dataframe als 'Container' für die Resultate definieren
articles <- data.frame()

# Mit einem Loop alle Artikelangaben aus den konvertierten Artikeln extrahieren
for (name in docnames) {
  ID <- gsub(''\\.txt'', '''', name)
  APS_shortname <- unlist(str_split(ID, ''\\.''))[1]
  doc <- read.table(paste0(directory, name), header = F, quote = '''', fileEncoding =
     'UTF-8', sep = '\t')
  text <- unlist(str_split('####', doc))[3]
  title <- unlist(str_split('####', doc))[2]
  meta <- unlist(str_split('####', doc))[1]
  pub_date <- unlist(str_split('','', meta))[2]
  page <- unlist(str_split('','', meta))[2]
  articles <- rbind(articles, cbind(APS_shortname, ID, text, pub_date, page, title))
}

# Identifikationsnummern für die autmomatische Vorcodierung und die endgültige Codierung
# hinzufügen (sind beide noch leer zu Beginn)
articles$issue_ID_final <- 0
articles$issue_ID_auto <- 0

# Identifikationsnummern für Coders und Zeitungen aus den vorher aktualisierten Tabellen
# holen und im Artikeldatensatz ergänzen
articles$source_ID <- NA
articles$coder_ID <- NA
```

```
for(i in 1:length(articles[,1])) {
  SQL_source <- paste(''select source_ID from sources where APS_shortname='', articles[,1][i], ''
  articles$source_ID[i] <- as.numeric(as.character(dbGetQuery(con, SQL_source)))

  SQL_coder <- paste(''select coder_ID from sources where APS_shortname='', articles[,1][i], '''
  articles$coder_ID[i] <- as.numeric(as.character(dbGetQuery(con, SQL_coder)))
}

# Zeitungstabelle ergänzen
dbWriteTable(con, ''articles'', articles, append = T, row.name = F)
```

Hier müssen zunächst die konvertierten Artikel eingelesen werden. Danach können aus den bis dahin aktualisierten Tabellen die Identifikationsnummern der Coder und Zeitungen hinzugefügt werden, bevor die Artikeltabelle aktualisiert wird. Nach der Speicherung der Artikel in der SQL-Datenbank können diese falls vorgesehen durch eine automatisierte Vorklassifikation bearbeitet werden und danach über eine Codierapplikation fertig codiert werden. Abbildung 2.1.3 zeigt einen Screenshot des Browser-Frontends des Prototyps der Codierapplikation. Die Applikation ist im Ordner *APS-Shiny* des Dropbox-Verzeichnisses enthalten. Sie kann direkt aus *R* wie im Skript *ui.R* beschrieben gestartet werden, indem der Ordner *APS-Shiny.* welcher die Applikationsskripte enthält, mit dem Befehl *runApp()* aufgerufen wird. Diesem Prototyp fehlen noch die Anbindungen an die Datenbank, aber die Eingabefelder und das laden von Zeitungsartikeln funktionieren. Für eine lokale Anwendung wird gleichzeitig ein Browser mit dem Frontend gestartet, in einer alltagstauglichen Version müsste die *Shiny*-Applikation auf einem Webserver gestartet werden, um die gleichzeitige Codierung durch mehrere Coder zu erlauben.

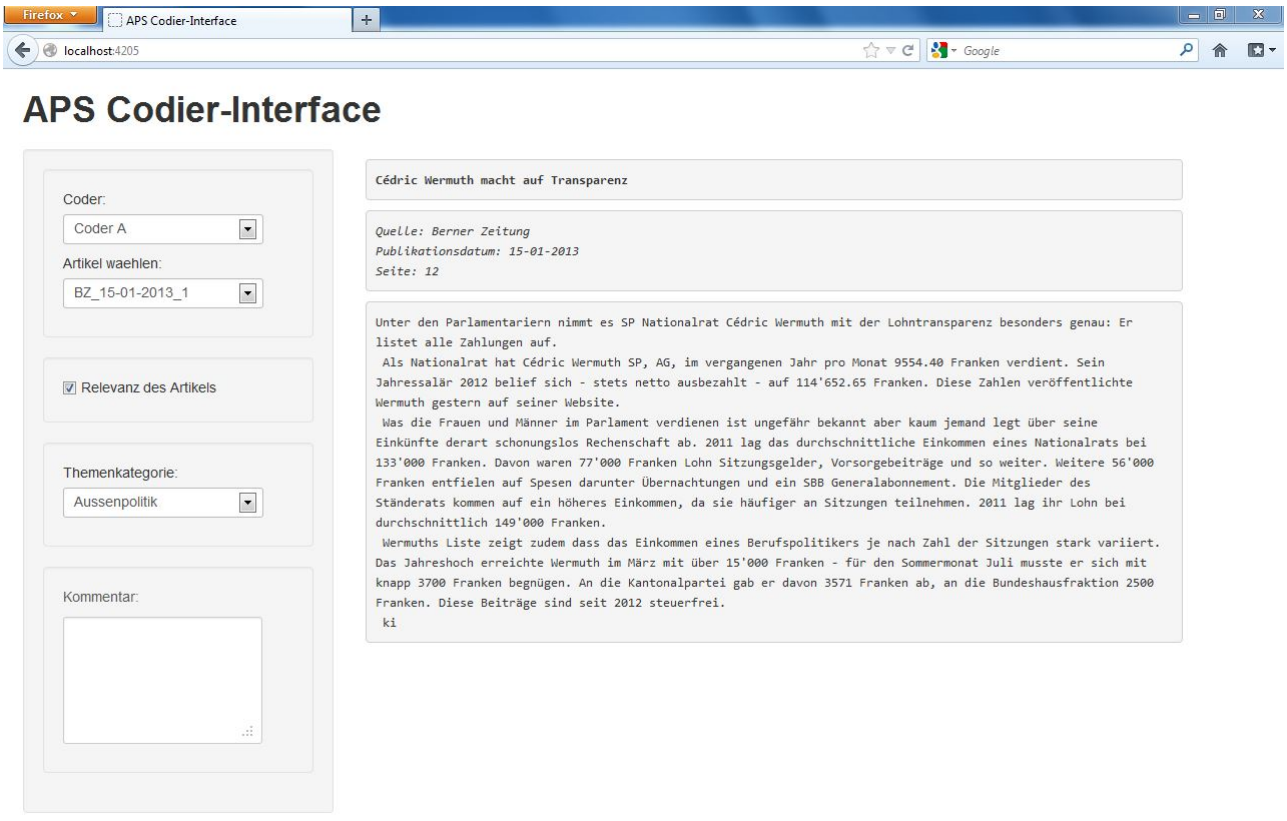

Abbildung 8: Browser-Frontend der Shiny-Codierapplikation

Rechts sind die Anzeigepanels für den Text und die Metadaten des zu bearbeitetenden Artikels vorgesehen. Links sind die Engabefelder angeordnet. Die Codierenden sollen zunächst aus der Liste der Ihnen zugeordneten Artikel einen Zeitungsartikel auswählen können. Danach sollen, falls vorhanden, Vorschläge der autmoatisierten Vorselektion und Vorklassifikation bereits in den Feldern zur Relevanz und der Themenkategorie angezeigt werden – hier ist der Artikel als relevant und zugehörig zur Kategorie *Auusenpolitik* angezeigt. Schliesslich ist noch die Angabemöglichkeit für Kommentare vorgesehen. Zusammenfassen kann dieser Prototyp bereits aufzeigen, dass eine effiziente und von verschiedenen Computern über den Browser zugängliche Webappliaktion, welche die APS-Zeitungsdokumentation systematisieren helfen würde, mit nicht sehr grossem Aufwand aufgebaut werden kann.

2.2 Qualitätssicherung

Vergleichbar zu der für die Pilot-Studie durchgeführten Messung der Qualität der automatisierten Erhebung möchte dieser Bericht einfache aber effektive Verfahren vorschlagen, wie die Reliabilität[10] der Dokumentation systematisch und kontinuierlich geprüft und ausgewiesen werden kann. Eine Qualitätssicherung ist aus der Sicht des Verfassers für eine Zeitungsdokumentation in der Grösse und Detailliertheit, wie sie beim APS erfolgt, von zentraler Bedeutung, ungeachtet ob das APS in Zukunft Anstrengungen in Richtung Automatisierung unternimmt oder bei der bewährten Art der manuellen Dokumentation bleibt. Mit systematischen Angaben über die Reliabilität kann das APS seinem Anspruch einer lückenlosen Dokumentation des politischen Geschehens in der Schweiz gerecht werden und der wissenschaftlichen Qualität des Jahrbuchs Nachdruck verleihen. In desem Zusammenhang muss noch erwähnt werden, dass der Autor dieses Berichts während der Pilot-Studie festgestellt hat, dass die Qualität der APS-Zeitungsdokumentation zumindest für das bearbeitete Jahr 2013 von sehr hoher Qualität ist. Für den automatisierten Download der relevanten Zeitungsartikel zum Beispiel war es unbedingt notwendig, dass die Titel, Zeitungs- und Datumsangaben konsistent eingetragen wurden, was in der absoluten Mehrheit der Fall war (in genau 99% der Fälle bzw. 9'921 von 10'000 Artikeln).

Es gibt in der wissenschaftlichen Literatur (Jurafsky and Martin, 2000; Krippendorff, 2004; Neuendorf, 2002; Lombard, Snyder-Dutch and Bracken, 2002) eine übereinstimmende Meinung zu einigen grundlegenden Richtlinien, wie eine solche Qualitätskontrolle erfolgen soll:

1. Die Qualität sollte an einer Zufallsstichprobe von 10% aller klassifizierten Texte geprüft werden.
2. Wichtig ist die Messung der Intercoder-Reliabilität (die Übereinstimmung zwischen den an der Erhebung beteiligten Codierenden und der eingesetzten Verfahren) sowie die Intracoder-Reliabilität (Übereinstimmung der Codierenden und Verfahren über die Zeit).
3. Die verwendeten Messwerte müssen die Risiken durch *false positives* (falsch zugeteilte Fälle) und *false negatives* (nicht zugeteilte Fälle) berücksichtigen.

Aufgrund dieser Richtlinien ist ein konkretes Szenario vorstellbar, wie die Qualität der APS-Dokumentation gesichert werden kann:

1. Jeder zehnte eintreffende Artikel wird als relevant für die Reliabilitätstest eingestuft.
2. Dieser Artikel wird von allen beteiligten Personen erfasst. Dabei soll möglichst verhindert werden, dass die Personen wissen, dass es sich um einen Testartikel handelt.
3. Der gleiche Artikel wird von der zuständigen Person nach 2 Monaten nochmals erfasst.

[10] Reliabilität bedeutet eine Qualitätsangabe, inwiefern an der Messung beteiligte Personen oder ein Messinstrument wirklich dasjenige messen, was beabsichtigt war. Im vorliegenden Fall geht es darum, dass überprüft werden kann, inwiefern das APS-Klassifikationsschema während der Dokumentationsarbeit wirklich umgesetzt wird.

4. Damit entsteht ein Korpus von mehrfach codierten Zeitungsartikeln, welcher systematisch mit den in der Pilot-Studie vorgestellten Kennzahlen (*Recall* und *Precision*) nach Stärken und Schwächen untersucht werden kann. Beide Kennzahlen werden im Vergleich zu einer Referenzcodierung berechnet. Dies sollte immer die erste Codierung derjenigen Person sein, welche für einen Artikel eigentlich zuständig ist (d.h. der/die Experte/in für die Zeitung oder das Thema). Normalerweise wird davon ausgegangen, dass Werte über 0.8 als gut erachtet werden. Eine solche fixe Grenze kann aber trügerisch sein. Für einfache Klassifizierungen, z.B. ob ein Zeitungsartikel überhaupot relevant ist für das APS-Archiv, sollte ein höherer Wert gesetzt werden (z.B. 0.9) als für relativ komplexe Klassifizierungen wie z.B. auf der feinsten Ebene des APS-Klassifizierungsschemas, wo zwischen sich sehr stark überschneidenden Themen unterschieden werden muss. Hier kann eine tiefere Schwelle (z.B. 0.7 oder gar 0.6) angesetzt werden.

5. Wenn die beiden Kennzahlen unter einen bestimmten Wert fallen, kann mittels genauerer Codieranweisungen oder Optimierung der Klassifikationssoftware eine Verbesserung angestrebt werden.

3 Zusammenfassung und Perspektiven

Die hier beschriebene Pilostudie für eine potenzielle Automatisierung der APS-Zeitungsdokumentation hat die folgenden Haupterkenntnisse zu Tage gefördert. Der erste Teil beschrieb eine systematische Evaluation bestehender automatischer Klassifierungsverfahren. Diese Evaluation wurde an einer Auswahl von 9'921 Zeitungsartikeln für die inhaltlichen Klassifikationen bzw. 19'909 Artikeln für die Selektion sowie mit 192 verschiedenen Klassifikationsläufen durchgeführt, was höchsten Standards der Textanalyse entspricht. Die wichtigsten Resultate sind, dass a) die Erkennung der Metadaten vollständig automatisiert werden könnte, b) die Selektion mit einer Reliabilität erfolgen kann, welche nur noch eine manuelle Nachkontrolle einer kleinen Stichprobe erfordert, und c) die inhaltliche Klassifikation zwar eine systematischen Nachkontrolle erfordert, aber sehr viel Effizienz einbringt. Entgegen den Erwartungen sind zudem Selektionen und Klassifikationen auf Ebene der einzelenen Zeitungen und feineren Kategorienschemata in der Tendenz sogwar besser als Klassifikationen über die Gesamtzahl der Artikel. Im zweiten Teil wurden mögliche Änderungen der Prozessabläufe diskutiert und aufgezeigt, dass zumindest ein Teil des möglichen Effizienzgewinnes in die systematische Überprüfung der Relibailität der APS-zeitungsdokumentation investiert werden könnte. Zudem wurde ein Prototyp einer Codierapplikation vorgestellt, welche die APS-Zeitungsdokumentation unterstützen könnte, unabhängig davon ob eine Automatisierung geplant wird oder nicht.

Literatur

APS. 2013. *Elektronische Zeitungsdokumentation – Handbuch.* Bern, CH: Année Politique Suisse / Schweizer Jahrbuch der Politik.

Feinerer, I., K. Hornik and D. Meyer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25(5):1–54.

Friedman, J., T. Hastie and R. Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software* 33(1):1.

Jurafsky, D and J. H. Martin. 2000. *Speech an Langauge Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice-Hall.

Jurka, T. P. 2012. "maxent: An R package for low-memory multinomial logistic regression with support for semi-automated text classification." *The R Journal* 4(1):56–59.

Jurka, T. P., L. Collingwood, A. E. Boydstun, E. Grossman and W. van Atteveldt. 2013. "RTextTools: A Supervised Learning Package for Text Classification." *The R Journal* 5(1):6–12.

Krippendorff, Klaus. 2004. *Content Analysis. An Introduction to Its Methodology.* London, UK: Sage Publications.

Lombard, M., J. Snyder-Dutch and C. C. Bracken. 2002. "Content Analysis in Mass communication: Assessment and Reporting of Intercoder Reliability." *Human Communication Research* 28(4):587–604.

Meyer, D. 2012. *Support Vector Machines.* Technische Universität Wien, Austria.

Neuendorf, K. A. 2002. *The Content Analysis Guidebook.* Thousand Oaks, CA: Sage Publications.